Variational Bayesian Inference for Audio-Visual Tracking of Multiple Speakers

Yutong Ban, Xavier Alameda-Pineda, Laurent Girin, Radu Horaud

Variational Bayesian Inference for Audio-Visual Tracking of Multiple Speakers

Yutong Ban, Xavier Alameda-Pineda, *IEEE Senior Member*, Laurent Girin and Radu Horaud

***Abstract*—In this paper we address the problem of tracking multiple speakers via the fusion of visual and auditory information. We propose to exploit the complementary nature and roles of these two modalities in order to accurately estimate smooth trajectories of the tracked persons, to deal with the partial or total absence of one of the modalities over short periods of time, and to estimate the acoustic status – either speaking or silent – of each tracked person over time. We propose to cast the problem at hand into a generative audio-visual fusion (or association) model formulated as a latent-variable temporal graphical model. This may well be viewed as the problem of maximizing the posterior joint distribution of a set of continuous and discrete latent variables given the past and current observations, which is intractable. We propose a variational inference model which amounts approximating the joint distribution with a factorized distribution. The solution takes the form of a closed-form expectation maximization procedure. We describe in detail the inference algorithm, we evaluate its performance and we compare it with several baseline methods. These experiments show that the proposed audio-visual tracker performs well in informal meetings involving a time-varying number of people.**

***Index Terms*—Audio-visual tracking, multiple object tracking, dynamic Bayesian networks, variational inference, expectation-maximization, speaker diarization.**

I. Introduction

In this paper we address the problem of tracking multiple speakers via the fusion of visual and auditory information [1]–[7]. We propose to exploit the complementary nature of these two modalities in order to accurately estimate the position of each person at each time step, to deal with the partial or total absence of one of the modalities over short periods of time, and to estimate the acoustic status, either speaking or silent, of each tracked person. We propose to cast the problem at hand into a generative audio-visual fusion (or association) model formulated as a latent-variable temporal graphical model. We propose a tractable solver via a variational approximation.

We are particularly interested in tracking people involved in informal meetings and social gatherings, e.g. Fig. 1. In this type of scenarios, participants wander around, cross each other, move in and out the camera field of view, take speech turns, etc. Acoustic room conditions, e.g. reverberation, and overlapping audio sources of various kinds drastically deteriorate or modify the microphone signals. Likewise, occluded persons, lighting conditions and mid-range camera distance complicate the task of visual processing. It is therefore impossible to gather reliable and continuous flows of visual **and** audio observations. Hence one must design a fusion and tracking method that is able to deal with intermittent visual and audio data.

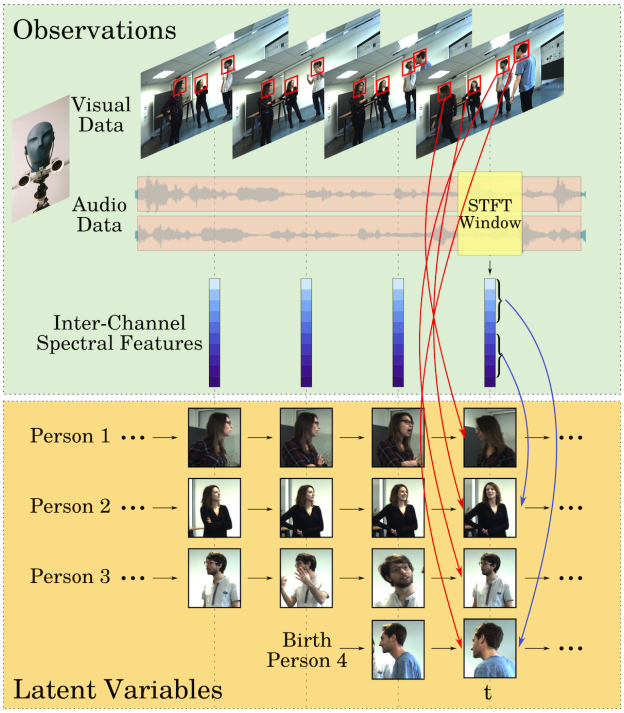


Fig. 1: Multiple speaker tracking is cast into the framework of Bayesian inference. Visual observations (person detections) and audio observations (inter-channel spectral features) are assigned to continuous latent variables (i.e. speaker positions) via discrete latent variables (one for each observation). As shown, the algorithm is causal (it uses only past and present observations) and incorporates a birth process to account for not yet seen/heard persons.

We propose a multi-speaker tracking method based on a dynamic Bayesian model that fuses audio and visual information over time from their respective observation spaces. This may well be viewed as a generalization of single-observation and single-target Kalman filtering – which yields an exact recursive solution – to multiple observations and multiple targets, which makes the exact recursive solution computationally intractable. We propose a variational approximation of the joint posterior distribution over the continuous variables (positions and velocities of tracked persons) and discrete variables (observation-to-person associations) at each time step, given all the past and

Y. Ban, X. Alameda-Pineda and R. Horaud are with Inria Grenoble Rhône-Alpes, Montbonnot Saint-Martin, France. L. Girin is with GIPSA Lab, Univ. Grenoble Alpes, France.

Funding from the European Union via the FP7 ERC Advanced Grant VHIA #340113 is greatly acknowledged.

present audio and visual observations. The proposed approximation consists on factorizing the joint distribution. We obtain a variational expectation maximisation (VEM) algorithm that is not only computationally tractable, but also very efficient.

In general, multiple object tracking consists of the temporal estimation of the kinematic state of each object, i.e. position and velocity. In computer vision, local descriptors are used to better discriminate between objects, e.g. person detectors/descriptors based on hand-crafted features [8] or on deep neural networks [9]. If the tracked objects emit sounds, their states can be inferred as well using sound-source localization techniques combined with tracking, e.g. [10]. These techniques are often based on the estimation of the sound's direction of arrival (DOA) using a microphone array, e.g. [11], or on a steered beamformer [10]. DOA estimation can be carried out either in the temporal domain [12], or in the spectral (Fourier) domain [13]. However, spectral-domain DOA estimation methods are more robust than temporal-domain methods, in particular in the presence of background noise and reverberation [14], [15]. The multiple sound-source localization and tracking method of [10] combines a steered beamformer with a particle filter. The loudest sound source is detected first, the second loudest one is next detected, etc., and up to four sources. This leads to many false detections. Particle filtering is combined with source-to-track assignment probabilities in order to determine whether a newly detected source is a false detection, a source that is currently being tracked, or a new source. In practice, this method requires several empirically defined thresholds.

Via proper camera-microphone calibration, audio and visual observations can be aligned such that a DOA corresponds to a 2D location in the image plane. In this paper we adopt the audio-visual alignment method of [16], which learns a mapping from the space spanned by *inter-channel spectral features* (audio features) to the space of source locations, which in our case corresponds to the image plane. Interestingly, the method of [16] estimates both this mapping and its inverse via a closed-form EM algorithm. Moreover, this allows us to exploit the richness of representing acoustic signals in the short-time Fourier domain [17] and to extract noise- and reverberation-free audio features [14].

We propose to represent the audio-visual fusion problem via two sets of independent variables, i.e. visual-feature-to-person and audio-feature-to-person sets of assignment variables. An interesting characteristic of this way of doing is that the proposed tracking algorithm can choose to use visual features, audio features, or a combination of both, and this choice can be made independently for every person and for every time step. Indeed, audio and visual information are rarely available simultaneously and continuously. Visual information suffers from limited camera field-of-view, occlusions, false positives, missed detections, etc. Audio information is often corrupted by room acoustics, environmental noise and overlapping acoustic signals. In particular speech signals are sparse, non-stationary and are emitted intermittently, with silence intervals between speech utterances. Hence a robust audio-visual tracking must explicitly take into account the temporal sparsity of the two modalities and this is exactly what is proposed in this paper.

We use the AV16.3 [18] and the AVDIAR [19] datasets to evaluate the performance of the proposed audio-visual tracker. We use the Multiple Object Tracking (MOT) metrics and the Optimal Sub Pattern Assignment fo Tracks (OSPA-T) metrics to quantitatively assess method performance. MOT and in particular MOTA (tracking accuracy), which combines false positives, false negatives, identity switches, by comparing the estimated tracks with the ground-truth trajectories, is a commonly used score to assess the quality of a multiple person tracker.[1] OSPA-T measures the distance between two point sets and hence it is also useful to compare ground-truth tracks with estimated tracks in the context of multi-target tracking [20]. We use MOT and OSPA-T metrics to compare our method with two recently proposed audio-visual tracking methods [4], [7] and with a visual tracker [8]. An interesting outcome of the proposed method is that speaker diarization, i.e. who speaks when, can be coarsely inferred from the tracking output, thanks to the audio-feature-to-person assignment variables. The speaker diarization results obtained with our method are compared with two other methods [19], [21] based on the Diarization Error Rate (DER) score.

The remainder of the paper is organized as follows. Section II describes the related work. Section III describes in detail the proposed formulation. Section IV describes the proposed variational approximation and Section V details the variational expectation-maximization procedure. The algorithm implementation is described in Section VI. Tracking results and comparisons with other methods are reported in Section VII. Finally, Section VIII draws a few conclusions.[2]

II. Related Work

In computer vision, there is a long history of multiple object tracking methods. While these methods provide interesting insights concerning the problem at hand, a detailed account of existing visual trackers is beyond the scope of this paper. Several audio-visual tracking methods were proposed in the recent past, e.g. [1]–[3], [22]. These papers proposed to use approximate inference of the filtering distribution using Markov chain Monte Carlo particle filter sampling (MCMC-PF). These methods cannot provide estimates of the accuracy and merit of each modality with respect to each tracked person.

More recently, audio-visual trackers based on particle filtering and probability hypothesis density (PHD) filters were proposed, e.g. [4]–[7], [23]–[25]. [6] used DOAs of audio sources to guide the propagation of particles, and combined the filter with a mean-shift algorithm to reduce the computational complexity. Some PHD filter variants were proposed to improve the tracking performance [23], [24]. The method of [4] also used DOAs of active audio sources to give more importance to particles located around DOAs. Along the same

[1] https://motchallenge.net/

[2] Supplemental materials are available at https://team.inria.fr/perception/research/var-av-track/

line of thought, [7] proposed a mean-shift sequential Monte Carlo PHD (SMC-PHD) algorithm that used audio information to improve the performance of a visual tracker. This implies that the persons being tracked must emit acoustic signals continuously and that multiple-source audio localization is reliable enough for proper audio-visual alignment.

PHD-based tracking methods are computationally efficient but their inherent limitation is that they are unable to associate observations to tracks. Hence they require an external post-processing mechanism that provides associations. Also, in the case of PF-based audio-visual filtering, the number of tracked persons must be set in advance and sampling can be a computational burden. In contrast, the proposed variational formulation embeds association variables within the model, uses a birth process to estimate the initial number of persons and to add new ones along time, and an explicit dynamic model yields smooth trajectories.

Another limitation of the methods proposed in [1], [3], [6], [23]–[25] is that they need as input a continuous flow of audio and visual observations. To some extent, this is also the case with [4], [7], where only the audio observations are supposed to be continuous. All these methods showed good performance in the case of the AV16.3 dataset [18] in which the participants spoke simultaneously and continuously – which is somehow artificial. The AV16.3 dataset was recorded in a specially equipped meeting room using three cameras that generally guarantee that frontal views of the participants were always available. This contrasts with the AVDIAR dataset which was recorded with one sensor unit composed of two cameras and six microphones. The AVDIAR scenarios are composed of participants that take speech turns while they look at each other, hence they speak intermittently and they do not always face the cameras.

Recently, we proposed an audio-visual clustering method [26] and an audio-visual speaker diarization method [19]. The weighted-data clustering method of [26] analyzed a short time window composed of several audio and visual frames and hence it was assumed that the speakers were static within such temporal windows. Binaural audio features were mapped onto the image plane and were clustered with nearby visual features. There was no dynamic model that allowed to track speakers. The audio-visual diarization method [19] used an external multi-object visual tracker that provided trajectories for each tracked person. The audio-feature-space to image-plane mapping [16] was used to assign audio information to each tracked person at each time step. Diarization itself was modeled with a binary state variable (speaking or silent) associated with each person. The diarization transition probabilities (state dynamics) were hand crafted, with the assumption that the speaking status of a person was independent of all the other persons. Because of the small number of state configurations, i.e. $\{0,1\}^N$ (where N is the maximum number of tracked persons), the MAP solution could be found by exhaustively searching the state space. In Section VII-I we use the AVDIAR recordings to compare our diarization results with the results obtained with [19].

The variational Bayesian inference method proposed in this paper may well be viewed as a multimodal generalization of variational expectation maximization algorithms for multiple object tracking using either visual-only information [8] or audio-only information [27], [28]. We show that these models can be extended to deal with observations living in completely different mathematical spaces. Indeed, we show that two (or several) different data-processing pipelines can be embedded and treated on an equal footing in the proposed formulation. Special attention is given to audio-visual alignment and to audio-to-person assignments: (i) we learn a mapping from the space of audio features to the image plane, as well as the inverse of this mapping, which are integrated in the proposed generative approach, and (ii) we show that the an increase in the number of assignment variables, due to the use of two modalities, do not affect the complexity of the algorithm. Absence of observed data of any kind or erroneous data are carefully modeled: this enables the algorithm to deal with intermittent observations, whether audio, visual, or both. This is probably one of the most prominent features of the method, in contrast with most existing audio-visual tracking methods which require continuous and simultaneous flows of visual and audio data.

This paper is an extended version of [29] and of [30]. The probabilistic model and its variational approximation were briefly presented in [29] together with preliminary results obtained with three AVDIAR sequences. Reverberation-free audio features were used in [30] where it was shown that good performance could be obtained with these features when the audio mapping was trained in one room and tested in another room. With respect to these two papers. we provide detailed descriptions of the proposed formulation, of the variational expectation maximization solver and of the implemented algorithm. We explain in detail the birth process, which is crucial for track initialization and for detecting potentially new tracks at each time step. We experiment with the entire AVDIAR dataset and we several sequences from the AV16.3 dataset; we benchmark our method with the state-of-the-art multiple-speaker audio-visual tracking methods [4], [7] and with [8]. Moreover, we show that our tracker can be used for audio-visual speaker diarization [19].

III. Proposed Model

A. Mathematical Definitions and Notations

Unless otherwise specified, uppercase letters denote random variables while lowercase letters denote their realizations, e.g. $p(X = x)$, where $p(\cdot)$ denotes either a probability density function (pdf) or a probability mass function (pmf). For the sake of conciseness we generally write $p(x)$. Vectors are written in slanted bold, e.g. $\boldsymbol{X}, \boldsymbol{x}$, whereas matrices are written in bold, e.g. $\mathbf{Y}, \mathbf{y}$. Video and audio data are assumed to be synchronized, and let t denote the common frame index. Let N be the upper bound of the number of persons that can simultaneously be tracked at any time t, and let $n \in \{1 \dots N\}$ be the person index. Let $n = 0$ denote

nobody. A t subscript denotes variable concatenation at time t, e.g. $\mathbf{X}_t = (\boldsymbol{X}_{t1}, \ldots, \boldsymbol{X}_{tn}, \ldots, \boldsymbol{X}_{tN})$, and the subscript $1:t$ denotes concatenation from 1 to t, e.g. $\mathbf{X}_{1:t} = (\mathbf{X}_1, \ldots, \mathbf{X}_t)$.

Let $\boldsymbol{X}_{tn} \in \mathcal{X} \subset \mathbb{R}^2$, $\boldsymbol{Y}_{tn} \in \mathcal{Y} \subset \mathbb{R}^2$ and $\boldsymbol{W}_{tn} \in \mathcal{W} \subset \mathbb{R}^2$ be three latent variables that correspond to the 2D position, 2D velocity and 2D size (width and height) of person n at t, respectively. Typically, $\boldsymbol{X}_{tn}$ and $\boldsymbol{W}_{tn}$ are the center and size of a bounding box of a person while $\boldsymbol{Y}_{tn}$ is the velocity of $\boldsymbol{X}_{tn}$. Let $\boldsymbol{S}_t = \{(\boldsymbol{X}_{tn}^\top, \boldsymbol{W}_{tn}^\top, \boldsymbol{Y}_{tn}^\top)^\top\}_{n=1}^N \subset \mathbb{R}^6$ be the complete set of continuous latent variables at t, where $^\top$ denotes the transpose operator. Without loss of generality, in this paper a person is characterized with the bounding box of her/his head and the center of this bounding box is assumed to be the location of the corresponding speech source.

We now define the observations. At each time t there are M_t visual observations and K_t audio observations. Let $\mathbf{f}_t = \{\boldsymbol{f}_{tm}\}_{m=1}^{M_t}$ and $\mathbf{g}_t = \{\boldsymbol{g}_{tk}\}_{k=1}^{K_t}$ be realizations of the visual and audio observed random variables $\{\boldsymbol{F}_{tm}\}_{m=1}^{M_t}$ and $\{\boldsymbol{G}_{tk}\}_{k=1}^{K_t}$, respectively. Visual observations, $\boldsymbol{f}_{tm} = (\boldsymbol{v}_{tm}^\top, \boldsymbol{u}_{tm}^\top)^\top$, correspond to the bounding boxes of detected faces, namely the concatenation of the bounding-box center, width and height, $\boldsymbol{v}_{tm} \in \mathcal{V} \subset \mathbb{R}^4$, and of a feature vector $\boldsymbol{u}_{tm} \in \mathcal{H} \subset \mathbb{R}^d$ that describes the photometric content of that bounding box, i.e. a d-dimensional face descriptor (Section VII-D). Audio observations, $\boldsymbol{g}_{tk}$, correspond to inter-channel spectral features, where k is a frequency sub-band index. Let's assume that there are K sub-bands, that $K_t \leq K$ sub-bands are *active* at t, i.e. sub-bands with sufficient signal energy, and that there are J frequencies per sub-band. Hence, $\boldsymbol{g}_{tk} \in \mathbb{R}^{2J}$ corresponds to the real and imaginary parts of J complex-valued Fourier coefficients. It is well established that inter-channel spectral features $\{\boldsymbol{g}_{tk}\}_{k=1}^{K_t}$ contain audio-source localization information, which is what is needed for tracking. These audio features are obtained by applying the multi-channel audio processing method described in Section VII-C below. Note that both the number of visual and of audio observations at t, M_t and K_t, vary over time. Let $\mathbf{o}_{1:t} = (\mathbf{o}_1, \ldots, \mathbf{o}_t)$ denote the set of observations from 1 to t, where $\mathbf{o}_t = (\mathbf{f}_t, \mathbf{g}_t)$.

Finally, we define the assignment variables of the proposed latent variable model. There is an assignment variable (a discrete random variable) associated with each observed variable. Namely, let A_{tm} and B_{tk} be associated with $\boldsymbol{f}_{tm}$ and with $\boldsymbol{g}_{tk}$, respectively, e.g. $p(A_{tm} = n)$ denotes the probability of assigning visual observation m at t to person n. Note that $p(A_{tm} = 0)$ and $p(B_{tk} = 0)$ are the probabilities of assigning visual observation m and audio observation k to none of the persons, or to nobody. In the visual domain, this may correspond to a false detection while in the audio domain this may correspond to an audio signal that is not uttered by a person. There is an additional assignment variable, C_{tk} that is associated with the audio generative model described in Section III-D. The assignment variables are jointly denoted with $\boldsymbol{Z}_t = (\boldsymbol{A}_t, \boldsymbol{B}_t, \boldsymbol{C}_t)$.

B. The Filtering Distribution

We remind that the objective is to estimate the positions and velocities of participants (multiple person tracking) and, possibly, to estimate their speaking status (speaker diarization). The audio-visual multiple-person tracking problem is cast into the problems of estimating the filtering distribution $p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t})$ and of inferring the state variable $\mathbf{S}_t$. Subsequently, speaker diarization can be obtained from audio-feature-to-person information via the estimation of the assignment variables $\boldsymbol{B}_{tk}$ (Section VI-C).

We reasonably assume that the state variable $\mathbf{S}_t$ follows a first-order Markov model, and that the visual and audio observations only depend on $\mathbf{S}_t$ and $\boldsymbol{Z}_t$. By applying Bayes rule, one can then write the filtering distribution of $(\mathbf{s}_t, \boldsymbol{z}_t)$ as:

$$p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t}) \propto p(\mathbf{o}_t|\mathbf{s}_t, \boldsymbol{z}_t)p(\boldsymbol{z}_t|\mathbf{s}_t)p(\mathbf{s}_t|\mathbf{o}_{1:t-1}), \quad (1)$$

with:

$$p(\mathbf{o}_t|\mathbf{s}_t, \boldsymbol{z}_t) = p(\mathbf{f}_t|\mathbf{s}_t, \boldsymbol{a}_t)p(\mathbf{g}_t|\mathbf{s}_t, \boldsymbol{b}_t, \boldsymbol{c}_t), \quad (2)$$

$$p(\boldsymbol{z}_t|\mathbf{s}_t) = p(\boldsymbol{a}_t)p(\boldsymbol{b}_t)p(\boldsymbol{c}_t|\boldsymbol{s}_t, \boldsymbol{b}_t), \quad (3)$$

$$p(\mathbf{s}_t|\mathbf{o}_{1:t-1}) = \int p(\mathbf{s}_t|\mathbf{s}_{t-1})p(\mathbf{s}_{t-1}|\mathbf{o}_{1:t-1})d\mathbf{s}_{t-1}. \quad (4)$$

Eq. (2) is the joint (audio-visual) observed-data likelihood. Visual and audio observations are assumed independent conditionally to $\mathbf{S}_t$, and their distributions will be detailed in Sections III-C and III-D, respectively.[3] Eq. (3) is the prior distribution of the assignment variable. The observation-to-person assignments are assumed to be a priori independent so that the probabilities in (3) factorize as:

$$p(\boldsymbol{a}_t) = \prod_{m=1}^{M_t} p(a_{tm}), \quad (5)$$

$$p(\boldsymbol{b}_t) = \prod_{k=1}^{K_t} p(b_{tk}), \quad (6)$$

$$p(\boldsymbol{c}_t|\mathbf{s}_t, \boldsymbol{b}_t) = \prod_{k=1}^{K_t} p(c_{tk}|\boldsymbol{s}_{tn}, B_{tk} = n). \quad (7)$$

It makes sense to assume that these distributions do not depend on t and that they are uniform. The following notations are introduced: $\eta_{mn} = p(A_{tm} = n) = 1/(N+1)$ and $\rho_{kn} = p(B_{tk} = n) = 1/(N+1)$. The probability $p(c_{tk}|\boldsymbol{s}_{tn}, B_{tk} = n)$ is discussed below (Section III-D).

Eq. (4) is the predictive distribution of $\mathbf{s}_t$ given the past observations, i.e. from 1 to $t-1$. The state dynamics in (4) are modeled with a linear-Gaussian first-order Markov process. Moreover, it is assumed that the dynamics are independent over speakers:

$$p(\mathbf{s}_t|\mathbf{s}_{t-1}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{s}_{tn}; \mathbf{D}\boldsymbol{s}_{t-1\,n}, \boldsymbol{\Lambda}_{tn}), \quad (8)$$

[3]We will see that $\mathbf{G}_t$ depends on $\mathbf{X}_t$ but depends neither on $\mathbf{W}_t$ nor on $\mathbf{Y}_t$, and $\mathbf{F}_t$ depends on $\mathbf{X}_t$ and $\mathbf{W}_t$ but not on $\mathbf{Y}_t$.

where $\mathbf{\Lambda}_{tn}$ is the dynamics' covariance matrix and $\mathbf{D}$ is the state transition matrix, given by:

$$\mathbf{D} = \begin{pmatrix} \mathbf{I}_{4\times4} & \mathbf{I}_{2\times2} \\ \mathbf{0}_{2\times4} & \mathbf{I}_{2\times2} \end{pmatrix}.$$

As described in Section IV below, an important feature of the proposed model is that the predictive distribution (4) at frame t is computed from the state dynamics model (8) and an approximation of the filtering distribution $p(\mathbf{s}_{t-1}|\mathbf{o}_{1:t-1})$ at frame $t-1$, which also factorizes across speaker. As a result, the computation of (4) factorizes across speakers as well.

C. The Visual Observation Model

As already mentioned above (Section III-A), a visual observation $\boldsymbol{f}_{tm}$ consists of the center, width and height of a bounding box, namely $\boldsymbol{v}_{tm} \in \mathcal{V} \subset \mathbb{R}^4$, as well as of a feature vector $\boldsymbol{u}_{tm} \in \mathcal{H} \subset \mathbb{R}^d$ describing the region inside the bounding box. Since the velocity is not observed, a 4×6 projection matrix $\mathbf{P}_f = (\mathbf{I}_{4\times4}\ \mathbf{0}_{4\times2})$ is used to project $\boldsymbol{s}_{tn}$ onto $\mathcal{V}$. Assuming that the M_t visual observations $\{\boldsymbol{f}_{tm}\}_{m=1}^{M_t}$ available at t are independent, and that the appearance representation of a person is independent of his/her position in the image, e.g. CNN-based embedding, the visual likelihood in (2) is defined as:

$$p(\mathbf{f}_t|\mathbf{s}_t, \boldsymbol{a}_t) = \prod_{m=1}^{M_t} p(\boldsymbol{v}_{tm}|\mathbf{s}_t, a_{tm}) p(\boldsymbol{u}_{tm}|\mathbf{h}, a_{tm}), \quad (9)$$

where the observed bounding-box centers, widths, heights, and feature vectors are drawn from the following distributions:

$$p(\boldsymbol{v}_{tm}|\mathbf{s}_t, A_{tm} = n) = \begin{cases} \mathcal{N}(\boldsymbol{v}_{tm}; \mathbf{P}_f \boldsymbol{s}_{tn}, \mathbf{\Phi}_{tm}) & \text{if } 1 \le n \le N \\ \mathcal{U}(\boldsymbol{v}_{tm}; \text{vol}(\mathcal{V})) & \text{if } n = 0, \end{cases} \quad (10)$$

$$p(\boldsymbol{u}_{tm}|\mathbf{h}, A_{tm} = n) = \begin{cases} \mathcal{B}(\boldsymbol{u}_{tm}; \boldsymbol{h}_n) & \text{if } 1 \le n \le N \\ \mathcal{U}(\boldsymbol{u}_{tm}; \text{vol}(\mathcal{H})) & \text{if } n = 0, \end{cases} \quad (11)$$

where $\mathbf{\Phi}_{tm} \in \mathbb{R}^{4\times4}$ is a covariance matrix quantifying the measurement error in the bounding-box center and size, $\mathcal{U}(\cdot; \text{vol}(\cdot))$ is the uniform distribution with $\text{vol}(\cdot)$ being the volume of the support of the variable, $\mathcal{B}(\cdot; \boldsymbol{h})$ is the Bhattacharya distribution [31], and $\mathbf{h} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N) \in \mathbb{R}^{d\times N}$ is a set of prototype feature vectors that model the appearances of the N persons.

D. The Audio Observation Model

It is well established in the recent audio signal processing literature that inter-channel spectral features encode sound-source localization information [13], [14], [16]. Therefore, observed audio features, $\mathbf{g}_t = \{\boldsymbol{g}_{tk}\}_{k=1}^{K_t}$ are obtained by considering all the pairs of a microphone array. Audio observations depend neither on the size of the bounding box $\boldsymbol{w}_t$, nor on the velocity $\boldsymbol{y}_t$. Indeed, we note that the velocity of a sound source (a moving person) is of about 1 meter/second, which is negligible compared to the speed of sound. Moreover, the inter-microphone distance is small compared to the source-to-microphone distance, hence the Doppler effect, if any, is similar across microphones. Hence one can replace $\boldsymbol{s}$ with $\boldsymbol{x} = \mathbf{P}_g \boldsymbol{s}$ in the equations below, with $\mathbf{P}_g = (\mathbf{I}_{2\times2}\ \mathbf{0}_{2\times4})$. By assuming independence across frequency sub-bands (indexed by k), the audio likelihood in (2) can be factorized as:

$$p(\mathbf{g}_t|\mathbf{s}_t, \boldsymbol{b}_t, \boldsymbol{c}_t) = \prod_{k=1}^{K_t} p(\boldsymbol{g}_{tk}|\boldsymbol{x}_{tb_{tk}}, b_{tk}, c_{tk}). \quad (12)$$

While the inter-channel spectral features $\boldsymbol{g}_{tk}$ contain localization information, in complex acoustic environments there is no explicit transformation that maps a source location onto an inter-channel spectral feature. We therefore make recourse to modeling this mapping via learning a non-linear regression. We use the method of [14] to extract audio features and the piecewise-linear regression model of [32] to learn a mapping between the space of audio-source locations and the space of audio features. The method of [32] belongs to the mixture of experts (MOE) class of models and hence it embeds well in our latent-variable mixture model. Let $\{h_{kr}\}_{r=1}^{r=R}$ be a set of linear regressions, such that the r-th linear transformation h_{kr} maps $\boldsymbol{x} \in \mathbb{R}^2$ onto $\boldsymbol{g}_k \in \mathbb{R}^{2J}$ for the frequency sub-band k. It follows that (12) writes:

$$p(\boldsymbol{g}_{tk}|\boldsymbol{x}_{tn}, B_{tk} = n, C_{tk} = r) = \begin{cases} \mathcal{N}(\boldsymbol{g}_{tk}; h_{kr}(\boldsymbol{x}_{tn}), \mathbf{\Sigma}_{kr}) & \text{if } 1 \le n \le N \\ \mathcal{U}(\boldsymbol{g}_{tk}; \text{vol}(\mathcal{G})) & \text{if } n = 0, \end{cases} \quad (13)$$

where $\mathbf{\Sigma}_{kr} \in \mathbb{R}^{2J\times2J}$ is a covariance matrix that captures the linear-mapping error and C_{tk} is a discrete random variable, such that $C_{tk} = r$ means that the audio feature $\boldsymbol{g}_{tk}$ is generated through the r-th linear transformation. Please consult Appendix A for details on how the parameters of the linear transformations h_{kr} are learned from a training dataset.

IV. Variational Approximation

Direct estimation of the filtering distribution $p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t})$ is computationally intractable. Consequently, evaluating expectations over this distribution is intractable as well. We overcome this problem via variational inference and associated EM closed-form solver [33], [34]. More precisely $p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t})$ is approximated with the following factorized form:

$$p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t}) \approx q(\mathbf{s}_t, \boldsymbol{z}_t) = q(\mathbf{s}_t) q(\boldsymbol{z}_t), \quad (14)$$

which implies

$$q(\mathbf{s}_t) = \prod_{n=1}^{N} q(\boldsymbol{s}_{tn}),\ q(\boldsymbol{z}_t) = \prod_{m=1}^{M_t} q(a_{tm}) \prod_{k=1}^{K} q(b_{tk}, c_{tk}), \quad (15)$$

where $q(A_{tm} = n)$ and $q(B_{tk} = n, C_{tk} = r)$ are the variational posterior probabilities of assigning visual observation m to person n and audio observation k to person

n, respectively. The proposed variational approximation (14) amounts to break the conditional dependence of $\mathbf{S}$ and $\boldsymbol{Z}$ with respect to $\mathbf{o}_{1:t}$ which causes the computational intractability. Note that the visual, $\boldsymbol{A}_t$, and audio, $\boldsymbol{B}_t$, $\boldsymbol{C}_t$, assignment variables are independent, that the assignment variables for each observation are also independent, and that B_{tk} and C_{tk} are conditionally dependent on the audio observation. This factorized approximation makes the calculation of $p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t})$ tractable. The optimal solution is given by an instance of the variational expectation maximization (VEM) algorithm [33], [34], which alternates between two steps:

- *Variational* E-step: the approximate log-posterior distribution of each one of the latent variables is estimated by taking the expectation of the complete-data log-likelihood over the remaining latent variables, i.e. (16), (17), and (18) below, and
- M-step: model parameters are estimated by maximizing the variational expected complete-data log-likelihood.[4]

In the case of the proposed model the latent variable log-posteriors write:

$$\log q(\boldsymbol{s}_{tn}) = \mathbb{E}_{q(\boldsymbol{z}_t)\prod_{\ell\neq n} q(\boldsymbol{s}_{t\ell})}[\log p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t})] + \text{const}, \tag{16}$$

$$\log q(a_{tm}) = \mathbb{E}_{q(\mathbf{s}_t)\prod_{\ell\neq m} q(a_{t\ell})\prod_k q(b_{tk},c_{tk})}[\log p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t})] + \text{const}, \tag{17}$$

$$\log q(b_{tk}, c_{tk}) = \mathbb{E}_{q(\mathbf{s}_t)\prod_m q(a_{tm})\prod_{\ell\neq k} q(b_{t\ell},c_{t\ell})}[\log p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t})] + \text{const}. \tag{18}$$

A remarkable consequence of the factorization (14) is that $p(\mathbf{s}_{t-1}|\mathbf{o}_{1:t-1})$ is replaced with $q(\mathbf{s}_{t-1}) = \prod_{n=1}^N q(\boldsymbol{s}_{t-1\,n})$, consequently (4) becomes:

$$p(\mathbf{s}_t|\mathbf{o}_{1:t-1}) \approx \int p(\mathbf{s}_t|\mathbf{s}_{t-1}) \prod_{n=1}^N q(\boldsymbol{s}_{t-1\,n}) d\mathbf{s}_{t-1}. \tag{19}$$

It is now assumed that the variational posterior distribution $q(\boldsymbol{s}_{t-1\;n})$ is Gaussian with mean $\boldsymbol{\mu}_{t-1\;n}$ and covariance $\boldsymbol{\Gamma}_{t-1\;n}$:

$$q(\boldsymbol{s}_{t-1\;n}) = \mathcal{N}(\boldsymbol{s}_{t-1\;n}; \boldsymbol{\mu}_{t-1\;n}, \boldsymbol{\Gamma}_{t-1\;n}). \tag{20}$$

By substituting (20) into (19) and combining it with (8), the predictive distribution (19) becomes:

$$p(\mathbf{s}_t|\mathbf{o}_{1:t-1}) \approx \prod_{n=1}^N \mathcal{N}(\boldsymbol{s}_{tn}; \mathbf{D}\boldsymbol{\mu}_{t-1\,n}, \mathbf{D}\boldsymbol{\Gamma}_{t-1\,n}\mathbf{D}^\top + \boldsymbol{\Lambda}_{tn}). \tag{21}$$

Note that the above distribution factorizes across persons. Now that all the factors in (1) have tractable expressions, a VEM algorithm can be derived.

[4]Even if the M-step is in closed-form, the inference is based on the variational posterior distributions. Therefore, the M-step could also be regarded as *variational*.

V. Variational Expectation Maximization

The proposed VEM algorithm iterates between an E-S-step, an E-Z-step, and an M-step on the following grounds.

1) E-S-step: the per-person variational posterior distribution of the state vector $q(\boldsymbol{s}_{tn})$ is evaluated by developing (16). The joint posterior $p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t})$ in (16) is the product of (2), (3) and (21). We thus first sum the logarithms of (2), of (3) and of (21). Then we ignore the terms that do not involve $\boldsymbol{s}_{tn}$. Evaluation of the expectation over all the latent variables except $\boldsymbol{s}_{tn}$ yields the following Gaussian distribution:

$$q(\boldsymbol{s}_{tn}) = \mathcal{N}(\boldsymbol{s}_{tn}; \boldsymbol{\mu}_{tn}, \boldsymbol{\Gamma}_{tn}), \tag{22}$$

with:

$$\boldsymbol{\Gamma}_{tn} = \Bigg(\underbrace{\sum_{k=1}^K \sum_{r=1}^R \beta_{tknr} \mathbf{P}_g^\top \mathbf{L}_{kr}^\top \boldsymbol{\Sigma}_{kr}^{-1} \mathbf{L}_{kr} \mathbf{P}_g}_{\#1} + \underbrace{\sum_{m=1}^{M_t} \alpha_{tmn} \mathbf{P}_f^\top \boldsymbol{\Phi}_{tm}^{-1} \mathbf{P}_f}_{\#2} + \underbrace{\left(\boldsymbol{\Lambda}_{tn} + \mathbf{D}\boldsymbol{\Gamma}_{t-1\,n}\mathbf{D}^\top\right)^{-1}}_{\#3} \Bigg)^{-1}, \tag{23}$$

and with:

$$\boldsymbol{\mu}_{tn} = \boldsymbol{\Gamma}_{tn} \Bigg(\underbrace{\sum_{k=1}^K \sum_{r=1}^R \beta_{tknr} \mathbf{P}_g^\top \mathbf{L}_{kr}^\top \boldsymbol{\Sigma}_{kr}^{-1} (\boldsymbol{g}_{kr} - \boldsymbol{l}_{kr})}_{\#1} + \underbrace{\sum_{m=1}^{M_t} \alpha_{tmn} \mathbf{P}_f^\top \boldsymbol{\Phi}_{tm}^{-1} \boldsymbol{v}_{tm}}_{\#2} + \underbrace{\left(\boldsymbol{\Lambda}_{tn} + \mathbf{D}\boldsymbol{\Gamma}_{t-1\,n}\mathbf{D}^\top\right)^{-1} \mathbf{D}\boldsymbol{\mu}_{t-1\,n}}_{\#3} \Bigg), \tag{24}$$

where $\alpha_{tmn} = q(A_{tm} = n)$ and $\beta_{tknr} = q(B_{tk} = n, C_{tk} = r)$ are computed in the E-Z-step below. A key point is that, because of the recursive nature of the formulas above, it is sufficient to make the Gaussian assumption at $t = 1$, i.e. $q(\boldsymbol{s}_{1n}) = \mathcal{N}(\boldsymbol{s}_{1n}; \boldsymbol{\mu}_{1n}, \boldsymbol{\Gamma}_{1n})$, whose parameters may be easily initialized. It follows that $q(\boldsymbol{s}_{tn})$ is Gaussian at every frame.

We note that both (23) and (24) are composed of three terms: the first (#1), second (#2) and third terms (#3) of (23) and of (24) correspond to the audio, visual, and past cumulated information contributions to the precision matrix and the mean vector, respectively. Remind that the covariance $\boldsymbol{\Phi}_{tm}$ is associated with the visual observed variable in (10). Matrices $\mathbf{L}_{kr}$ and vectors $\boldsymbol{l}_{kr}$ characterize the piecewise affine mappings from the space of person locations to the space of audio features, i.e. Appendix A, and covariances $\boldsymbol{\Sigma}_{kr}$ capture the errors that are associated with both audio measurements and the piecewise affine approximation in (13). A similar interpretation holds for the three terms of (24).

2) E-Z-step: by developing (17), along the same reasoning as above, we obtain the following closed-form expression for the variational posterior distribution of the visual assignment

variable:

$$\alpha_{tmn} = q(A_{tm} = n) = \frac{\tau_{tmn}\eta_{mn}}{\sum_{i=0}^{N}\tau_{tmi}\eta_{mi}}, \tag{25}$$

where τ_{tmn} is given by:

$$\tau_{tmn} = \begin{cases} \mathcal{N}(\boldsymbol{v}_{tm}; \mathbf{P}_f\boldsymbol{\mu}_{tn}, \boldsymbol{\Phi}_{tm})e^{-\frac{1}{2}\mathrm{tr}\left(\mathbf{P}_f^\top\boldsymbol{\Phi}_{tm}^{-1}\mathbf{P}_f\boldsymbol{\Gamma}_{tn}\right)} \\ \times\ \mathcal{B}(\boldsymbol{u}_{tm}; \boldsymbol{h}_n) & \text{if } 1 \leq n \leq N \\ \mathcal{U}(\boldsymbol{v}_{tm}; \mathrm{vol}(\mathcal{V}))\mathcal{U}(\boldsymbol{u}_{tm}; \mathrm{vol}(\mathcal{H})) & \text{if } n = 0. \end{cases}$$

Similarly, for the variational posterior distribution of the audio assignment variables, developing (18) leads to:

$$\beta_{tknr} = q(B_{tk} = n, C_{tk} = r) = \frac{\kappa_{tknr}\rho_{kn}\pi_r}{\sum_{i=0}^{N}\sum_{j=1}^{R}\kappa_{tkij}\rho_{ki}\pi_j}, \tag{26}$$

where κ_{tknr} is given by:

$$\kappa_{tknr} = \tag{27}$$
$$\begin{cases} \mathcal{N}(\boldsymbol{g}_{tk}; \mathbf{L}_{kr}\mathbf{P}_g\boldsymbol{\mu}_{tn} + \boldsymbol{l}_{kr}, \boldsymbol{\Sigma}_{kr})e^{-\frac{1}{2}\mathrm{tr}\left(\mathbf{P}_g^\top\mathbf{L}_{kr}^\top\boldsymbol{\Sigma}_{kr}^{-1}\mathbf{L}_{kr}\mathbf{P}_g\boldsymbol{\Gamma}_{tn}\right)} \\ \times\ \mathcal{N}(\tilde{\boldsymbol{x}}_{tn}; \boldsymbol{\nu}_r, \boldsymbol{\Omega}_r)) & \text{if } 1 \leq n \leq N \\ \mathcal{U}(\boldsymbol{g}_{tk}; \mathrm{vol}(\mathcal{G})) & \text{if } n = 0. \end{cases}$$

To obtain (27), an additional approximation is made. Indeed, the logarithm of (39) in Appendix A is part of the complete-data log-likelihood and the denominator of this formula contains a weighted sum of Gaussian distributions. Taking the expectation of this term is not tractable because of the denominator. Based on the dynamical model (8), we replace the state variable $\boldsymbol{x}_{tn}$ in (39) with a "naive" estimate $\tilde{\boldsymbol{x}}_{tn}$ predicted from the position and velocity inferred at $t-1$: $\tilde{\boldsymbol{x}}_{tn} = \boldsymbol{x}_{t-1\,n} + \boldsymbol{y}_{t-1\,n}$.

3) M-step: The entries of the covariance matrix of the state dynamics, $\boldsymbol{\Lambda}_{tn}$, are the only parameters that need be estimated. To this aim, we develop $\mathbb{E}_{q(\mathbf{s}_t)q(\boldsymbol{z}_t)}[\log p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t})]$ and ignore the terms that do not depend on $\boldsymbol{\Lambda}_{tn}$. We obtain:

$$J(\boldsymbol{\Lambda}_{tn}) = \mathbb{E}_{q(\boldsymbol{s}_{tn})}\big[\log\mathcal{N}(\boldsymbol{s}_{tn}; \mathbf{D}\boldsymbol{\mu}_{t-1\,n}, \mathbf{D}\boldsymbol{\Gamma}_{t-1\,n}\mathbf{D}^\top + \boldsymbol{\Lambda}_{tn})\big],$$

which can be further developed as:

$$J(\boldsymbol{\Lambda}_{tn}) = \log|\mathbf{D}\boldsymbol{\Gamma}_{t-1\,n}\mathbf{D}^\top + \boldsymbol{\Lambda}_{tn}| + \mathrm{Tr}\big((\mathbf{D}\boldsymbol{\Gamma}_{t-1\,n}\mathbf{D}^\top + \boldsymbol{\Lambda}_{tn})^{-1} \times \big((\boldsymbol{\mu}_{tn} - \mathbf{D}\boldsymbol{\mu}_{t-1\,n})(\boldsymbol{\mu}_{tn} - \mathbf{D}\boldsymbol{\mu}_{t-1\,n})^\top + \boldsymbol{\Gamma}_{tn}\big)\big). \tag{28}$$

Hence, by differentiating (28) with respect to $\boldsymbol{\Lambda}_{tn}$ and equating to zero, we obtain:

$$\boldsymbol{\Lambda}_{tn} = \boldsymbol{\Gamma}_{tn} - \mathbf{D}\boldsymbol{\Gamma}_{t-1\,n}\mathbf{D}^\top + (\boldsymbol{\mu}_{tn} - \mathbf{D}\boldsymbol{\mu}_{t-1\,n})(\boldsymbol{\mu}_{tn} - \mathbf{D}\boldsymbol{\mu}_{t-1\,n})^\top. \tag{29}$$

VI. Algorithm Implementation

The VEM procedure above will be referred to as VAVIT which stands for *variational audio-visual tracking*, and pseudo-code is shown in Algorithm 1. In theory, the order in which the two expectation steps are executed is not important. In practice, the issue of initialization is crucial. In our case, it is more convenient to start with the E-Z step rather than with the E-S step because the former is easier to initialize than the latter (see below). We start by explaining how the algorithm is initialized at $t = 1$ and then how the E-Z-step is initialized at each iteration. Next, we explain in detail the birth process. An interesting feature of the proposed method is that it allows to estimate who speaks when (i.e. perform speaker diarization) which is explained in detail at the end of the section.

Algorithm 1: Variational audio-visual tracking (VAVIT).

Input: visual observations $\mathbf{f}_{1:t} = \{\mathbf{v}_{1:t}, \xi_{1:t}\}$;
audio observations $\mathbf{g}_{1:t}$;
Output: Parameters of $q(\mathbf{s}_{1:t})$: $\{\boldsymbol{\mu}_{1:t,n}, \boldsymbol{\Gamma}_{1:t,n}\}_{n=0}^{N}$ (the estimated position of each person n is given by the two first entries of $\boldsymbol{\mu}_{1:t,n}$);
Person speaking status for $1:t$
Initialization (see Section VI-A);
for $t = 1$ *to end* **do**
  Gather visual and audio observations at frame t;
  Perform voice activity detection;
  Initialization of E-Z step (see Section VI-A);
  for $iter = 1$ *to* N_{iter} **do**
    E-Z-step (vision):
    for $m \in \{1, ..., M_t\}$ **do**
      for $n \in \{0, ..., N_t\}$ **do**
        Evaluate $q(A_{tm} = n)$ with (25);
      end
    end
    E-Z-step (audio):
    for $k \in \{1, ..., K_t\}$ **do**
      for $n \in \{0, ..., N_t\}$ *and* $r \in \{1, ..., R\}$ **do**
        Evaluate $q(B_{tk} = n, C_{tk} = r)$ with (26) and (27);
      end
    end
    E-S-step:
    for $n \in \{1, ..., N_t\}$ **do**
      Evaluate $\boldsymbol{\Gamma}_{tn}$ and $\boldsymbol{\mu}_{tn}$ with (23) and (24);
    end
    M-step: Evaluate $\boldsymbol{\Lambda}_{tn}$ with (29);
  end
  Perform birth (see Section VI-B);
  Output the results;
end

A. Initialization

At $t = 1$ one must provide initial values for the parameters of the distributions (22), namely $\boldsymbol{\mu}_{1n}$ and $\boldsymbol{\Gamma}_{1n}$ for all $n \in \{1 \dots N\}$. These parameters are initialized as follows. The means are initialized at the image center and the covariances are given very large values, such that the variational distributions $q(\boldsymbol{s}_{1n})$ are non-informative. Once these parameters are initialized, they remain constant for a few frames, i.e. until the birth process is activated (see Section VI-B below).

As already mentioned, it is preferable to start with the E-Z-step than with the E-S-step because the initialization of the former is straightforward. Indeed, the E-S-step (Section V) requires current values for the posterior probabilities (25) and (27) which are estimated during the E-Z-step and which are both difficult to initialize. Conversely, the E-Z-step only requires current mean values, $\boldsymbol{\mu}_{tn}$, which can be easily initialized by using the model dynamics (8), namely $\boldsymbol{\mu}_{tn} = \mathbf{D}\boldsymbol{\mu}_{t-1n}$.

B. Birth Process

We now explain in detail the birth process, which is executed at the start of the tracking to initialize a latent variable for each detected person, as well as at any time t to detect new persons. The birth process considers B consecutive visual frames. At t, with $t > B$, we consider the set of visual observations assigned to $n = 0$ from $t - B$ to t, namely observations whose posteriors (25) are maximized for $n = 0$ (at initialization all the observations are in this case). We then build observation sequences from this set, namely sequences of the form $(\tilde{\boldsymbol{v}}_{m_{t-B}}, \ldots, \tilde{\boldsymbol{v}}_{m_t})_{\tilde{n}} \in \mathcal{B}$, where m_t indexes the set of observations at t assigned to $n = 0$ and $\tilde{n}$ indexes the set $\mathcal{B}$ of all such sequences. Notice that the birth process only uses the bounding-box center, width and size, $\boldsymbol{v}$, and that the descriptor $\boldsymbol{u}$ is not used. Hence the birth process is only based on the smoothness of an observed sequence of bounding boxes. Let's consider the marginal likelihood of a sequence $\tilde{n}$, namely:

$$\begin{aligned}\mathcal{L}_{\tilde{n}} &= p((\tilde{\boldsymbol{v}}_{m_{t-B}}, \ldots, \tilde{\boldsymbol{v}}_{m_t})_{\tilde{n}}) \\ &= \int \cdots \int p(\tilde{\boldsymbol{v}}_{m_{t-B}} | \boldsymbol{s}_{t-B\,\tilde{n}}) \ldots p(\tilde{\boldsymbol{v}}_{m_t} | \boldsymbol{s}_{t\,\tilde{n}}) \\ &\times p(\boldsymbol{s}_{t\,\tilde{n}} | \boldsymbol{s}_{t-1\,\tilde{n}}) \ldots p(\boldsymbol{s}_{t-B+1\,\tilde{n}} | \boldsymbol{s}_{t-B\,\tilde{n}}) p(\boldsymbol{s}_{t-B\,\tilde{n}}) d\boldsymbol{s}_{t-B:t\,\tilde{n}},\end{aligned} \tag{30}$$

where $\boldsymbol{s}_{t,\tilde{n}}$ is the latent variable already defined and $\tilde{n}$ indexes the set $\mathcal{B}$. All the probability distributions in (30) were already defined, namely (8) and (10), with the exception of $p(\boldsymbol{s}_{t-B,\tilde{n}})$. Without loss of generality, we can assume that the latter is a normal distribution centered at $\tilde{\boldsymbol{v}}_{m_t}$ and with a large covariance. Therefore, the evaluation of (30) yields a closed-form expression for $\mathcal{L}_{\tilde{n}}$. A sequence $\tilde{n}$ generated by a person is likely to be smooth and hence $\mathcal{L}_{\tilde{n}}$ is high, while for a non-smooth sequence the marginal likelihood is low. A newborn person is therefore created from a sequence of observations $\tilde{n}$ if $\mathcal{L}_{\tilde{n}} > \tau$, where τ is a user-defined parameter. As just mentioned, the birth process is executed to initialize persons as well as along time to add new persons. In practice, in (30) we set $B = 3$ and hence, from $t = 1$ to $t = 4$ all the observations are initially assigned to $n = 0$.

C. Speaker Diarization

Speaker diarization consists of assigning temporal segment of speech to persons [35]. We introduce a binary variable χ_{tn} such that $\chi_{tn} = 1$ if person n speaks at time t and $\chi_{tn} = 0$ otherwise. Traditionally, speaker diarization is based on the following assumptions. First, it is assumed that speech signals are sparse in the time-frequency domain. Second, it is assumed that each time-frequency point in such a spectrogram corresponds to a single speech source. Therefore, the proposed speaker diarization method is based on assigning time-frequency points to persons.

In the case of the proposed model, speaker diarization can be coarsely inferred from frequency sub-bands in the following way. The posterior probability that the speech signal available in the frequency sub-band k at frame t was uttered by person n, given the audio observation $\boldsymbol{g}_{tk}$, is:

$$p(B_{tk} = n | \boldsymbol{g}_{tk}) = \sum_{r=1}^{R} p(B_{tk} = n, C_{tk} = r | \boldsymbol{g}_{tk}), \tag{31}$$

where B_{tk} is the audio assignment variable and C_{tk} is the affine-mapping assignment variable defined in Section III-D and in Appendix A. Using the variational approximation (26), this probability becomes:

$$\begin{aligned}p(B_{tk} = n | \boldsymbol{g}_{tk}) &\approx \sum_{r=1}^{R} q(B_{tk} = n, C_{tk} = r) \\ &= \sum_{r=1}^{R} \beta_{tknr},\end{aligned} \tag{32}$$

and by accumulating probabilities over all the frequency sub-bands, we obtain the following formula:

$$\chi_{tn} = \begin{cases} 1 & \text{if } \frac{1}{K_t} \sum_{k=1}^{K_t} \sum_{r=1}^{R} \beta_{tknr} \geq \gamma \\ 0 & \text{otherwise,} \end{cases} \tag{33}$$

where γ is a user-defined threshold. Note that there is no dynamic model associated with diarization: χ_{tn} is estimated independently at each frame and for each person. More sophisticated diarization models can be found in [19], [36].

VII. Experiments

A. The AVDIAR Dataset

We used the AVDIAR[5] dataset [19] to evaluate the performance of the proposed audio-visual tracking method. This dataset is challenging in terms of audio-visual analysis. There are several participants involved in informal conversations while wandering around. They are in between two and four meters away from the audio-visual recording device. They take speech turns and often there are speech overlaps. They turn their faces away from the camera. The dataset is annotated as follows: The visual annotations comprise the centers, widths and heights of two bounding boxes for each person and in each video frame, a face bounding box and an upper-body bounding box. An identity (a number) is associated with each person through the entire dataset. The audio annotations comprise the speech status of each person over time (speaking or silent), with a minimum speech duration of 0.2 s. The audio source locations correspond to the centers of the face bounding boxes.

[5] https://team.inria.fr/perception/avdiar/

The dataset was recorded with a sensor composed of two cameras and six microphones, but only one camera is used in the experiments described below. The videos were recorded at 25 FPS. The frame resolution is of 1920×1200 pixels corresponding to a field of view of $97^\circ \times 80^\circ$. The microphone signals are sampled at 16 kHz. The dataset was recorded into two different rooms, *living-room* and *meeting-room*, e.g. Fig. 3 and Fig. 4. These two rooms have quite different lighting conditions and acoustic properties (size, presence of furniture, background noise, etc.). Altogether there are 18 sequences associated with living-room (26927 video frames) and 6 sequences with meeting-room (6031 video frames). Additionally, there are two training datasets, $\mathcal{T}_1$ and $\mathcal{T}_2$ (one for each room) that contain input-output pairs of multichannel audio features and audio-source locations that allow to estimate the parameters (37) using the method of [16]. This yields a mapping between source locations in the image plane, $\boldsymbol{x}$, and audio features, $\boldsymbol{g}$. Audio feature extraction is described in detail below.

One interesting characteristic of the proposed tracking is its flexibility in dealing only with visual data, only with audio data, or with visual and audio data. Moreover, the algorithm is able to automatically switch from unimodal (audio or visual) to multimodal (audio and visual). In order to quantitatively assess the performance and merits of each one of these variants we used two configurations:

- *Full camera field of view (FFOV)*: The entire horizontal field of view of the camera, i.e. 1920 pixels, or 97°, is being used, such that visual and audio observations, **if any**, are simultaneously available, and
- *Partial camera field of view (PFOV)*: The horizontal field of view is restricted to 768 pixels (or 49°) and there are two *blind* strips (576 pixels each) on its left- and right-hand sides; the *audio field of view* remains unchanged, 1920 pixels, or 97°.

The PFOV configuration allows us to test scenarios in which a participant may leave the camera field of view and still be heard. Notice that since ground-truth annotations are available for the full field of view, it is possible to assess the performance of the tracker using audio observations only, as well as to analyse the behavior of the tracker when it switches from audio-only tracking to audio-visual tracking.

B. The AV16.3 Dataset

We also used the twelve recordings of the AV16.3 dataset [18] to evaluate the proposed method and to compare it with [4] and with [7]. The dataset was recorded in a meeting room. The videos were recorded at 25 FPS with three cameras fixed on the room ceiling. The image resolution is of 288×360 pixels. The audio signals were recorded with two eight-microphone circular arrays, both placed onto a table top, and sampled at 16 kHz. In addition, the dataset comes with internal camera calibration parameters, as well as with external calibration parameters, namely camera-to-camera and microphone-array-to-camera calibration parameters. We note that the scenarios associated with AV16.3 are somehow artificial in the sense that *the participants speak simultaneously and continuously*. This stays in contrast with the AVDIAR recordings where people take speech turns in informal conversations.

C. Audio Features

In the case of AVDIAR, the STFT (short-time Fourier transform) [17] is applied to each microphone signal using a 16 ms Hann window (256 audio samples per window) and with an 8 ms shift between successive windows (50% overlap), leading to 128 frequency bins and to 125 audio FPS. Inter-microphone spectral features are then computed using [15]. These features – referred to in [15] as *direct-path relative transfer function (DP-RTF) features* – are robust against background noise and against reverberations, hence they do not depend on the acoustic properties of the recording room, as they encode the direct path from the audio source to the microphones. Nevertheless, they may depend on the orientation of the speaker's face. If the microphones are positioned behind a speaker, the direct-path sound wave (from the speaker to the microphones) propagates through the speaker's head, hence it is attenuated. This may have a negative impact on the direct-to-reverberation ratio. Here we assume that, altogether, this has a limited effect.

The audio features are averaged over five audio frames in order to be properly aligned with the video frames. The feature vector is then split into $K = 16$ sub-bands, each sub-band being composed of $J = 8$ frequencies; sub-bands with low energy are disregarded. This yields the set of audio observations at t, $\{\boldsymbol{g}_{tk}\}_{k=1}^{K_t}$, $K_t \leq K$ (see Section III-D and Appendix A). Interestingly, the computed inter-microphone DP-RTF features can be mapped onto the image plane and hence they can be used to estimate directions of arrival (DOAs). Please consult [16], [32] for more details.

Alternatively, one can compute DOAs explicitly from time differences of arrival (TDOAs) between the microphones of a microphone array, provided that the inter-microphone geometry is known. The disadvantage is that DOAs based on TDOAs assume free-field acoustic-wave propagation and hence they don't have a built-in reverberation model. Moreover, if the camera parameters are known and if the camera location (extrinsic parameters) is known in the coordinate frame of the microphone array, as is the case with the AV16.3 dataset, it is possible to project DOAs onto the image plane. We use the multiple-speaker DOA estimator of [37] as it provides accurate results for the AV16.3 sensor setup [18]. Let d_{tk} be the line corresponding to the projection of a DOA onto the image plane and let $\boldsymbol{x}_{tn}$ be the location of person n at time t. It is straightforward to determine the point $\hat{\boldsymbol{x}}_{tk} \in d_{tk}$ the closest to $\boldsymbol{x}_{tn}$, e.g. Fig. 2. Hence the inter-channel spectral features $\{\boldsymbol{g}_{tk}\}_{k=1}^{K_t}$ are replaced with $\{\hat{\boldsymbol{x}}_{tk}\}_{k=1}^{K_t}$ and (13) is replaced

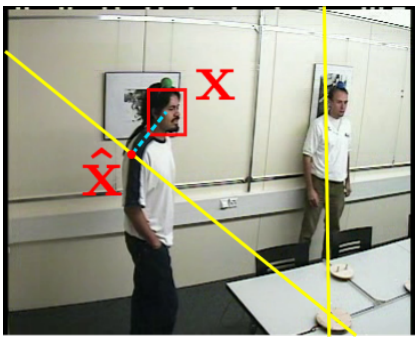


Fig. 2: This figure displays two DOAs, associated with one microphone array (bottom left), projected onto the image plane, and illustrates the geometric relationship between a DOA and the current location of a speaker.

with:

$$p(\hat{\boldsymbol{x}}_{tk}|\boldsymbol{x}_{tn}, B_{tk} = n) = \begin{cases} \mathcal{N}(\hat{\boldsymbol{x}}_{tk}; \boldsymbol{x}_{tn}, \sigma\mathbf{I}) & \text{if } 1 \leq n \leq N \\ \mathcal{U}(\hat{\boldsymbol{x}}_{tk}; \text{vol}(\mathcal{X})) & \text{if } n = 0, \end{cases} \tag{34}$$

where $\sigma\mathbf{I}$ is an isotropic covariance that models the uncertainty of the DOA, e.g. Fig. 5, third row.

D. Visual Features

In both AVDIAR and AV16.3 datasets participants do not always face the cameras and hence face detection is not robust. Instead we use the person detector of [38] from which we infer a body bounding-box and a head bounding-box. We use the person re-identification CNN-based method [39] to extract an embedding (i.e. a person descriptor) from the body bounding-box. This yields the feature vectors $\{\boldsymbol{u}_{tm}\}_{m=1}^{M_t} \subset \mathbb{R}^{2048}$ (Section III-C). Similarly, the center, width and height of the head bounding-box yield the observations $\{\boldsymbol{v}_{tm}\}_{m=1}^{M_t} \subset \mathbb{R}^4$ at each frame t.

E. Evaluation Metrics

We used standard multi-object tracking (MOT) metrics [40] to quantitatively evaluate the performance of the proposed tracking algorithm. The multi-object tracking accuracy (MOTA) is the most commonly used metric for MOT. It is a combination of false positives (FP), false negatives (FN; i.e. missed persons), and identity switches (IDs), and is defined as:

$$\text{MOTA} = 100\left(1 - \frac{\sum_t(\text{FP}_t + \text{FN}_t + \text{IDs}_t)}{\sum_t \text{GT}_t}\right), \tag{35}$$

where GT stands for the ground-truth person trajectories. After comparison with GT trajectories, each estimated trajectory can be classified as mostly tracked (MT) and mostly lost (ML) depending on whether a trajectory is covered by correct estimates more than 80% of the time (MT) or less than 20% of the time (ML). In the tables below, MT and ML indicated the percentage of ground-truth tracks under each situation.

TABLE I: OSPA-T and MOT scores for the living-room sequences (full camera field of view)

Method	OSPA-T(↓)	MOTA(↑)	FP(↓)	FN(↓)	IDs(↓)	MT(↑)	ML(↓)
[4]	28.12	10.37	44.64 %	43.95%	732	20%	7.5 %
[7]	30.03	18.96	8.13 %	72.09%	581	17.5%	52.5%
[8]	**14.79**	**96.32**	**1.77%**	**1.79%**	**80**	**92.5%**	**0%**
VAVIT	17.05	96.03	1.85%	2.0%	86	**92.5%**	**0%**

TABLE II: OSPA-T and MOT scores for the meeting-room sequences (full camera field of view).

Method	OSPA-T (↓)	MOTA(↑)	FP(↓)	FN(↓)	IDs(↓)	MT(↑)	ML(↓)
[4]	5.76	62.43	18.63%	17.19%	297	70.59 %	**0%**
[7]	7.83	28.48	0.93%	69.68%	155	0 %	52.94%
[8]	**3.02**	**98.50**	**0.25%**	**1.11%**	**25**	**100.00%**	**0%**
VAVIT	3.57	98.16	0.38%	1.27%	32	**100.00%**	**0%**

TABLE III: OSPA-T and MOT scores for the living-room sequences (partial camera field of view).

Method	OSPA-T(↓)	MOTA(↑)	FP(↓)	FN(↓)	IDs(↓)	MT(↑)	ML(↓)
[4]	28.14	17.82	36.86%	42.88%	1722	32.50%	7.5%
[7]	29.73	20.61	5.54%	72.45%	989	12.5%	40%
[8]	22.25	66.39	**0.48%**	32.95%	**129**	45%	7.5%
VAVIT	**21.77**	**69.62**	8.97%	**21.18%**	152	**70%**	**5%**

TABLE IV: OSPA-T and MOT scores for the meeting-room sequences (partial camera field of view).

Method	OSPA-T(↓)	MOTA(↑)	FP(↓)	FN(↓)	IDs(↓)	MT(↑)	ML(↓)
[4]	7.23	29.04	23.05%	45.19 %	461	29.41%	17.65%
[7]	8.17	26.95	1.05%	70.62%	234	5.88%	52.94%
[8]	**5.80**	64.24	**0.43%**	35.18%	**24**	36.84%	15.79%
VAVIT	5.81	**65.27**	5.07%	**29.5%**	26	**47.37%**	**10.53%**

In addition to MOT, we also used the OSPA-T metric [20]. OSPA-T is based on a distance between two point sets and combines various aspects of tracking performance, such as timeliness, track accuracy, continuity, data associations and false tracks. It should be noted that OSPA-T involves a number of parameters whose values must be provided in advance. We used the publicly available code provided by one of the authors of [20] for computing the OSPA-T scores in all the experimental evaluations reported below.[6]

In our experiments, the threshold of overlap to consider that a ground truth is covered by an estimation is set to 0.1 intersection over union (IoU). In the PFOV configuration, we need to evaluate the audio-only tracking, i.e. the speakers are in the blind areas. As mentioned before, audio localization is less accurate than visual localization. Therefore, for evaluating the audio-only tracker we relax by a factor of two the expected localization accuracy with respect to the audio-visual localization accuracy.

F. Benchmarking with Baseline Methods

To quantitatively evaluate its performance, we benchmarked the proposed method with two state-of-the-art audio-visual tracking methods. The first one is the audio-assisted video adaptive particle filtering (AS-VA-PF) method of [4], and

[6]http://ba-tuong.vo-au.com/codes.html

TABLE V: OSPA-T and MOT scores obtained with the AV16.3 dataset.

Method	OSPA-T (↓)	MOTA(↑)	FP(↓)	FN(↓)	IDs(↓)	MT(↑)	ML(↓)
[7]	17.28	36.4	16.72%	42.22%	765	11.11%	**0%**
[8]	13.32	82.9	**5.29%**	11.5 %	51	85.2%	**0%**
VAVIT	**10.88**	**84.1**	6.51%	**9.18%**	**29**	**92.6%**	**0%**

the second one is the sparse audio-visual mean-shift sequential Monte-Carlo probability hypothesis density (AV-MSSMC-PHD) method of [7]. Notice that both these methods do not make recourse to a person detector as they use a tracking-by-detection paradigm. This stays in contrast with our method which uses a person detector and probabilistically assigns each detection to each person. In principle, the baseline methods can be modified to accept person detection as visual information. However, we did not modify the baseline methods and used the software provided by the authors of [4] and [7]. Sound locations are used to reshape the typical Gaussian noise distribution of particles in a propagation step, then [4] uses the particles to weight the observation model. [7] uses audio information to improve the performance and robustness of a visual SMC-PHD filter. Both [4] and [7] require input from a multiple sound-source localization (SSL) algorithm. In the case of AVDIAR recordings, the multi-speaker localization method proposed in [15] is used to provide input to [4] and [7].[7] In the case of AV16.3 recordings the method of [18] is used to provide DOAs to [4], [7] and to our method, as explained in Section VII-C above.

We also compare the proposed method with a visual multiple-person tracker, more specifically the *online Bayesian variational tracker* (OBVT) of [8], which is based on a similar variational inference as the one presented in this paper. In [8] visual observations were provided by color histograms. In our benchmark, for the sake of fairness, the proposed tracker and [8] share the same visual observations, as described in Section VII-D.

The OSPA-T and MOT scores obtained with these methods as well as the proposed method are reported in Table I, Table II, Table III, Table IV, and Table V. The symbols ↑ and ↓ indicate higher the better and lower the better, respectively. In the case of AVDIAR, we report results with both meeting-room and living-room in the two configurations: FFOV, Table I and Table II and PFOV, Table III and Table IV. In the case of AV16.3 we report results with the twelve recordings commonly used by audio-visual tracking algorithms, Table V.

The most informative metrics are OSPA-T and MOTA (MOT accuracy) and one can easily see that both [8] and the proposed method outperform the other two methods. The poorer performance of both [4] and [7] for all the configurations is generally explained by the fact that these two methods expect audio and visual observations to be simultaneously available. In particular, [4] is not robust against visual occlusions, which leads to poor IDs (identity switches) scores.

The AV-MSSMC-PHD method [7] uses audio information in order to count the number of speakers. In practice, we noticed that the algorithm behaves differently with the two datasets. In the case of AVDIAR, we noticed that the algorithm assigns several visible participants to the same audio source, since in most of the cases there is only one active audio source at a time. In the case of AV16.3 the algorithm performs much better, since participants speak simultaneously and continuously. This explains why both FN (false negatives) and IDs (identity switches) scores are high in the case of AVDIAR, i.e. Tables I, II, and III.

One can notice that in the case of FFOV, [8] and the proposed method yield similar results in terms of OSPA-T and MOT scores: both methods exhibit low OSPA-T, FP, FN and IDs scores and, consequently, high MOTA scores. Moreover, they have very good MT and ML scores (out of 40 sequences 37 are mostly tracked, 3 are partially tracked, and none is mostly lost). As expected, the inferred trajectories are more accurate for visual tracking (whenever visual observations are available) than for audio-visual tracking: indeed, the latter fuses visual and audio observations which slightly degrades the accuracy because audio localization is less accurate than visual localization.

As for the PFOV configuration (Table III and Table IV), the proposed algorithm yields the best MOTA scores both for meeting-room and for living-room. Both [4] and [7] have difficulties when visual information is not available: both these algorithms fail to track speakers when they walk outside the visual field of view. While [7] can detect a speaker when it re-enters the visual field of view, [4] cannot. Obviously, the visual-only tracker [8] fails outside the camera field of view.

G. Audio-Visual Tracking Examples

We now provide and discuss results obtained with three AVDIAR recordings and one AV16.3 recording, namely the FFOV recording Seq13-4P-S2-M1 (Fig. 3), the PFOV recordings Seq19-2P-S1M1 (Fig. 4) and Seq22-1P-S0M1 (Fig. 6), and the seq45-3p-1111 recording of AV16.3 (Fig. 5).[8] All these recordings are challenging in terms of audio-visual tracking: participants are seated, then they stand up or they wander around. In the case of AVDIAR, some participants take speech turns and interrupt each other, while others remain silent.

The first rows of Fig. 3, Fig. 4 and Fig. 5 show four frames sampled from two AVDIAR recordings and one AV16.3 recording, respectively. The second rows show ellipses of constant density that correspond to visual uncertainty (covariances). The third rows show the audio uncertainty. The audio uncertainties (covariances) are much larger than the visual ones since audio localization is less accurate than visual localization. The fourth rows shows the contribution of the dynamic model to the uncertainty, i.e. the inverse of the precision (#3) in eq. (23). Notice that these "dynamic" covariances are small, in comparison with the "observation" covariances. This ensures tracking continuity (smooth tracjectories) when audio or visual observations are either weak or

[7]The authors of [4] and [7] kindly provided their software packages.

[8]https://team.inria.fr/perception/research/variational_av_tracking/

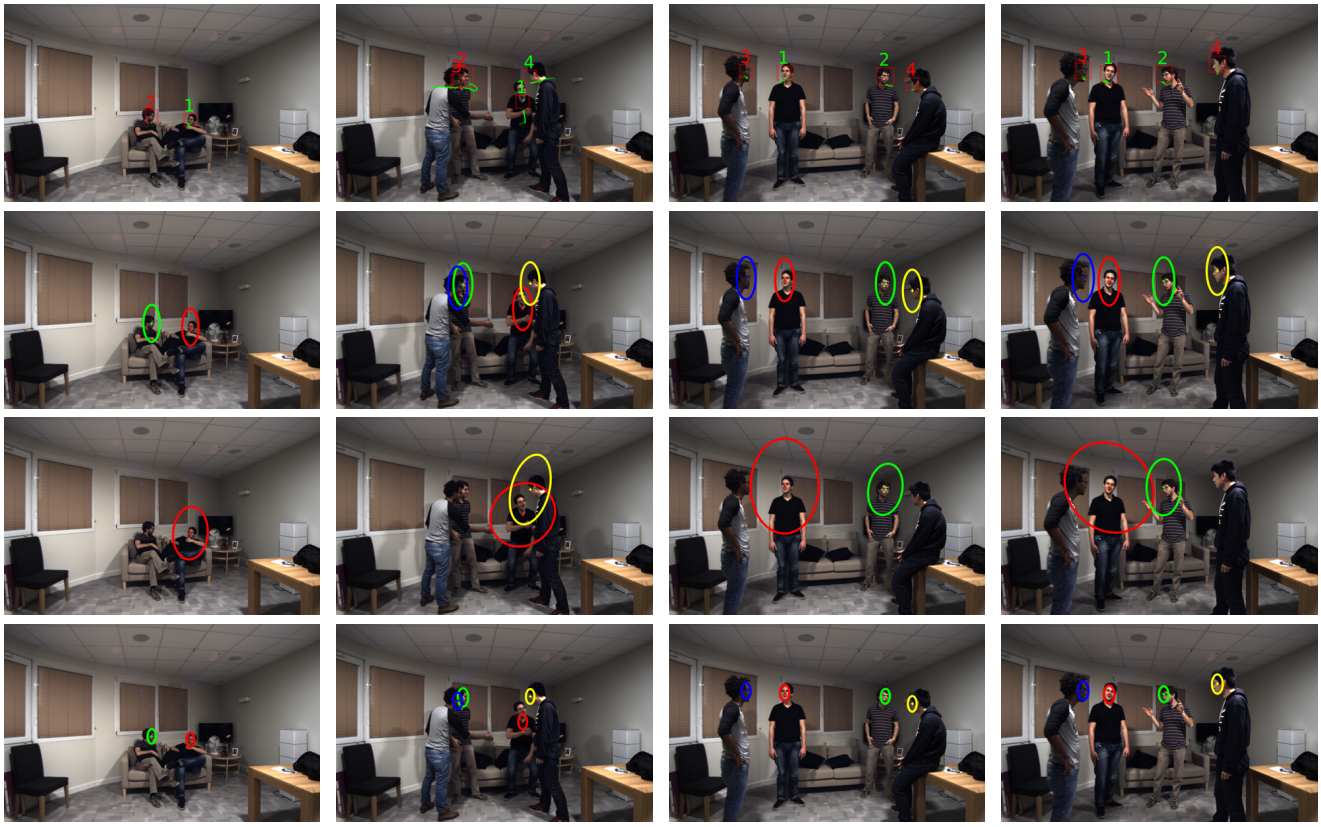


Fig. 3: Four frames sampled from Seq13-4P-S2M1 (living room). First row: green digits denote speakers while red digits denote silent participants. Second, third and fourth rows: the ellipses visualize the visual, audio, and dynamic covariances, respectively, of each tracked person. The tracked persons are color-coded: green, yellow, blue, and red.

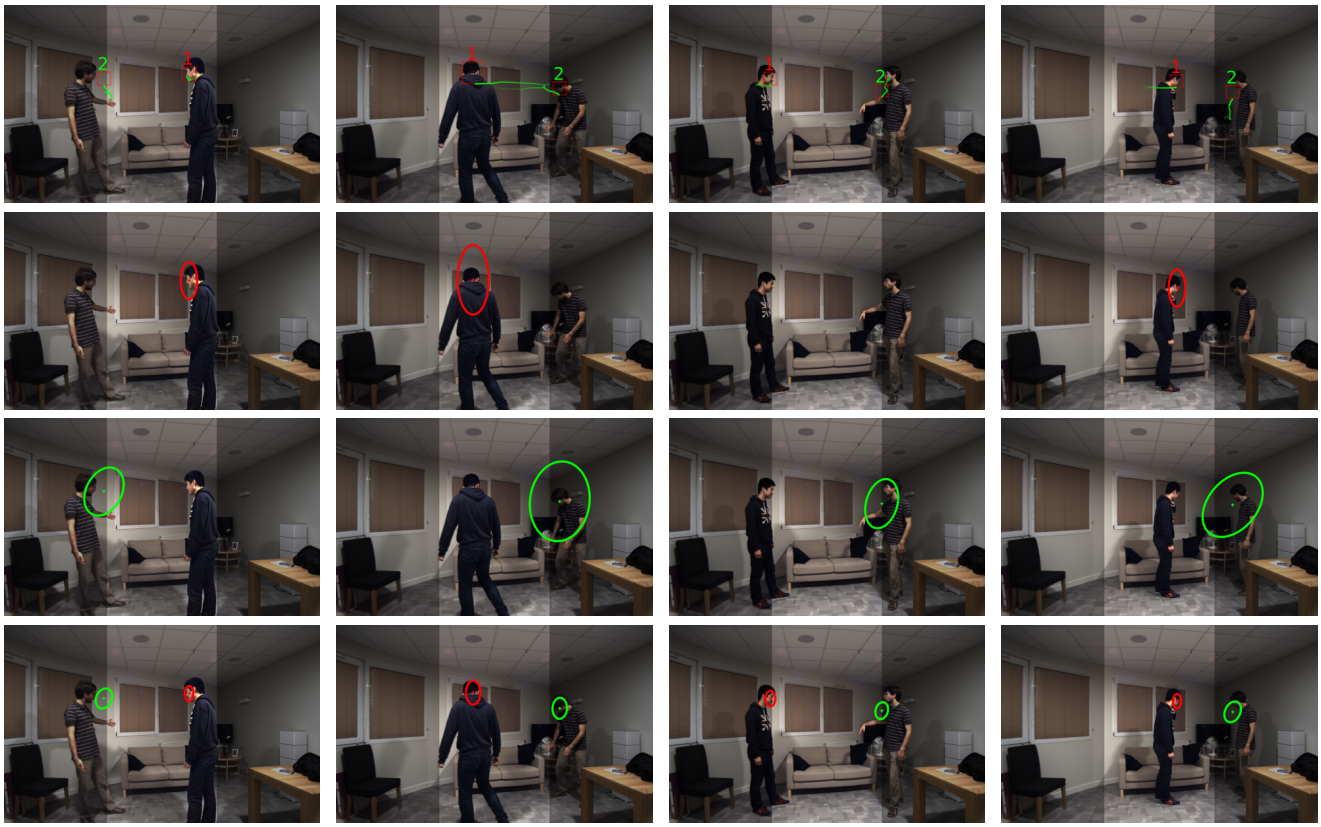


Fig. 4: Four frames sampled from Seq19-2P-S1M1 (living room). The camera field of view is limited to the central strip. Whenever the participants are outside the central strip, the tracker entirely relies on audio observations and on the model's dynamics.

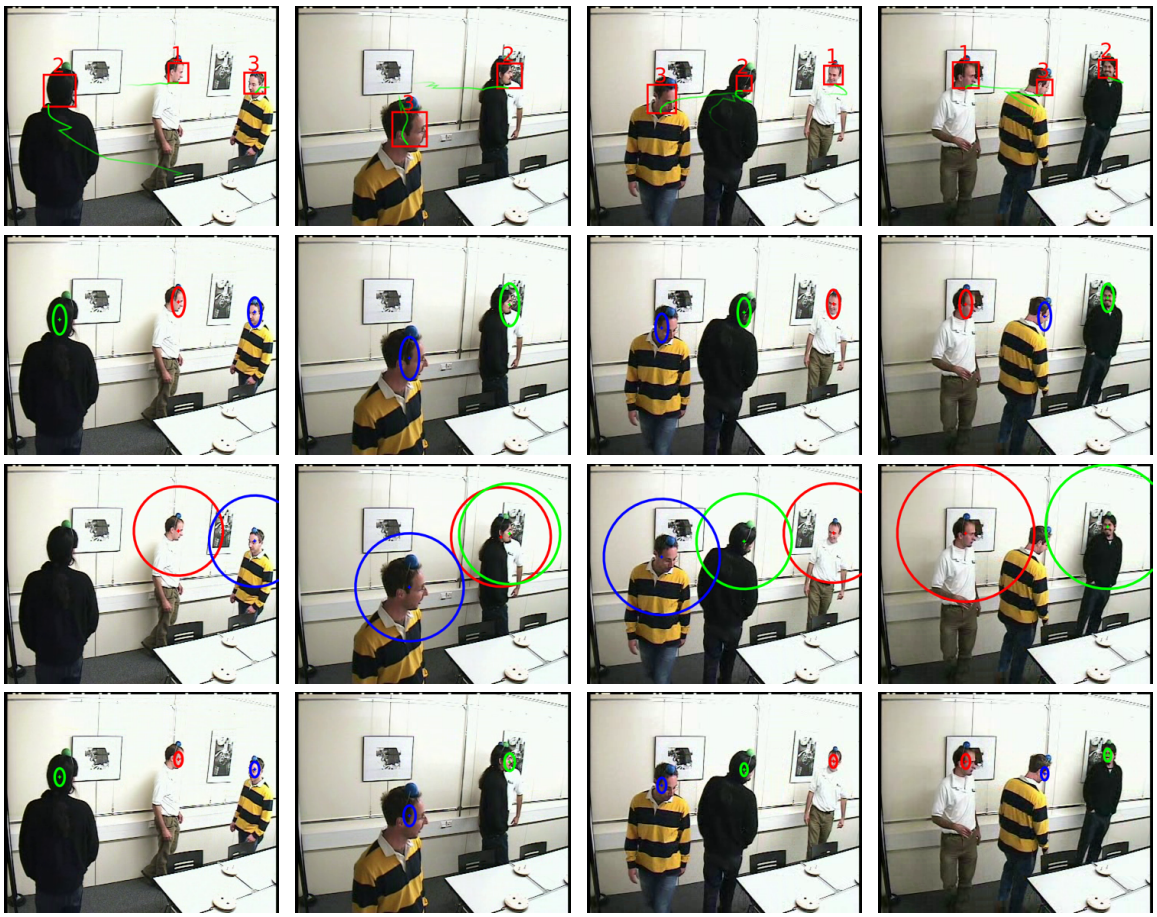
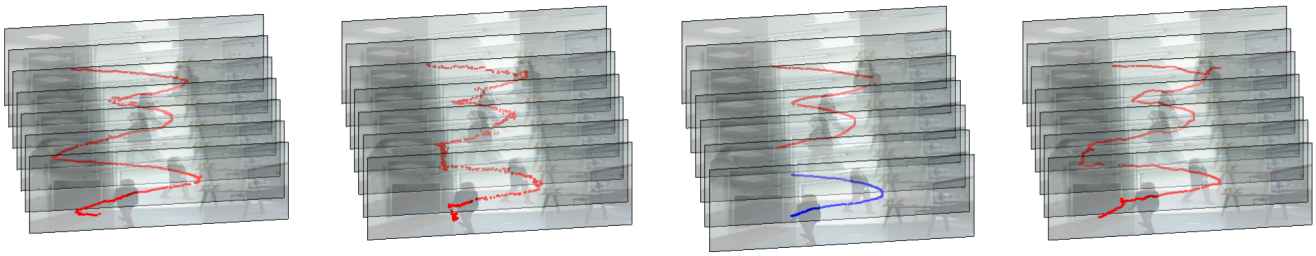


Fig. 5: Four frames sampled from seq45-3p-1111 of AV16.3. In this dataset, the participants speak simultaneously and continuously.



(a) Ground-truth trajectory (b) AS-VA-PF [4] (c) OBVT [8] (d) VAVIT (proposed)

Fig. 6: Trajectories associated with a tracked person under the PFOV configuration (sequence Seq22-1P-S0M1 recorded in meeting room). The ground-truth trajectory (a) corresponds to the center of the bounding-box of the head. The trajectory (b) obtained with [4] is non-smooth. Both [4] and [8] fail to track outside the camera field of view. In the case of the OBVT trajectory (c), there is an identity switch, from "red" (before the person leaves the visual field of view) to "blue" (after the person re-enters in the visual field of view).

totally absent. Fig. 4 shows a tracking example with a partial camera field of view (PFOV) configuration. In this case, audio and visual observations are barely available simultaneously. The independence of the visual and audio observation models and their fusion within the same dynamic model guarantees robust tracking in this case.

Fig. 6 shows the ground-truth trajectory of a person and the trajectories estimated with the audio-visual tracker [4], with the visual tracker [8], and with the proposed method. The ground-truth trajectory corresponds to a sequence of bounding-box centers. Both [4] and [8] failed to estimate a correct trajectory. Indeed, [4] requires simultaneous availability of audio-visual data while [8] cannot track outside the visual field of view. Notice the non-smooth trajectory obtained with [4] in comparison with the smooth trajectories obtained with variational inference, i.e. [8] and proposed.

H. Computation Times

Matlab implementations of algorithms [4], [7], [8] and VAVIT were run on an Intel(R) 8-core 2.40 GHz CPU E5-2609 equipped with 32 GB of RAM and with a GeForce GTX 1070

TABLE VI: Computation times (in seconds). All four algorithms are implemented in Matlab and run on the same computer.

Methods	AVDIAR: Living room	AVDIAR: Meeting room	AV16.3
Number of frames	26927	6031	11135
[4]	20821	2424	-
[7]	10510	2267	611
[8]	**542**	**130**	**236**
VAVIT	3456	759	260

GPU. The computation times provided in Table VI correspond to the total number of frames associated with all the sequences available in the two datasets. Both [8] and VAVIT necessitate a person detector. The CNN-based person detector runs on the same computer at 2 FPS. Person detection is run offline.

I. Speaker Diarization Results

As already mentioned in Section VI-C, speaker diarization information can be extracted from the output of the proposed VAVIT algorithm. Notice that, while audio diarization is an extremely well investigated topic, audio-visual diarization has received much less attention. In [36] it is proposed an audio-visual diarization method based on a dynamic Bayesian network that is applied to video conferencing. Their method assumes that participants take speech turns with a small silent interval between turns, which is an unrealistic hypothesis in the general case. The diarization method of [41] requires audio, depth and RGB data. More recently, [19] proposed a Bayesian dynamic model for audio-visual diarization that takes as input fused audio-visual information. Since diarization is not the main objective of this paper, we only compared our diarization results with [19], which achieves state of the art results, and with the diarization toolkit of [21] which only considers audio information.

The diarization error rate (DER) is generally used as a quantitative measure. As is the case with MOT, DER combines false positives (FP), false negatives (FN) and identity swithches (IDs) scores within a single metric. The NIST-RT evaluation toolbox[9] is used. The results obtained with [19], [21] and with the proposed method are reported in Table VII, for both the full field-of-view and partial field-of-view configurations (FFOV and PFOV). The proposed method performs better than the audio-only baseline method [21]. In comparison with [19], the proposed method performs slightly less well despite the lack of a special-purpose diarization model. Indeed, [19] implements diarization within a hidden Markov model (HMM) that takes into account both diarization dynamics and the audio activity observed at each time step, whereas our method is only based on observing the audio activity over time.

The ability of the proposed audio-visual tracker to perform diarization is illustrated in Fig. 7 and in Fig. 8 with a FFOV sequence (Seq13-4P-S2-M1, Fig. 3) and with a PFOV sequence (Seq19-2P-S1M1, Fig. 4), respectively.

[9]https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation

TABLE VII: DER (diarization error rate) scores obtained with the AVDIAR dataset.

Sequence	DiarTK [21]	[19]	Proposed (FFOV)	Proposed (PFOV)
Seq01-1P-S0M1	43.19	3.32	1.64	1.86
Seq02-1P-S0M1	49.9	-	2.38	2.09
Seq03-1P-S0M1	47.25	-	6.59	14.65
Seq04-1P-S0M1	32.62	9.44	4.96	10.45
Seq05-2P-S1M0	37.76	-	29.76	30.78
Seq06-2P-S1M0	56.12	-	14.72	15.83
Seq07-2P-S1M0	41.43	-	42.36	37.56
Seq08-3P-S1M1	31.5	-	38.4	48.86
Seq09-3P-S1M1	52.74	-	38.26	68.81
Seq10-3P-S1M1	56.95	-	54.26	54.04
Seq12-3P-S1M1	63.67	17.32	44.67	47.25
Seq13-4P-S2M1	47.56	29.62	43.45	43.17
Seq15-4P-S2M1	62.53	-	41.49	64.38
Seq17-2P-S1M1	17.24	-	16.53	15.63
Seq18-2P-S1M1	35.05	-	19.55	20.58
Seq19-2P-S1M1	38.96	-	26.47	27.84
Seq20-2P-S1M1	43.58	35.46	38.24	44.3
Seq21-2P-S1M1	32.22	20.93	25.87	25.9
Seq22-1P-S0M1	23.53	4.93	2.79	3.32
Seq27-3P-S1M1	46.05	18.72	47.07	54.75
Seq28-3P-S1M1	30.68	-	23.54	31.77
Seq29-3P-S1M0	38.68	-	30.74	35.92
Seq30-3P-S1M1	51.15	-	49.71	57.94
Seq32-4P-S1M1	41.51	30.20	46.25	43.03
Overall	42.58	**18.88**	28.73	33.36

VIII. Conclusions

We addressed the problem of tracking multiple speakers using audio and visual data. It is well known that the generalization of single-person tracking to multiple-person tracking is computationally intractable and a number of methods were proposed in the past. Among these methods, sampling methods based on particle filtering (PF) or on PHD filters have recently achieved the best tracking results. However, these methods have several drawbacks: (i) the quality of the approximation of the filtering distribution increases with the number of particles, which also increases the computational burden, (ii) the observation-to-person association problem is not explicitly modeled and a post-processing association mechanism must be invoked, and (iii) audio and visual observations must be available simultaneously and continuously. Some of these limitations were recently addressed both in [4] and in [7], where audio observations were used to compensate the temporal absence of visual observations. Nevertheless, people speak with pauses and hence audio observations are rarely continuously available.

In contrast, we proposed a variational approximation of the filtering distribution and we derived a closed-form variational expectation-maximization algorithm. The observation-to-person association problem is fully integrated in our model, rather than as a post-processing stage. The proposed VAVIT algorithm is able to deal with intermittent audio or visual observations, such that one modality can compensate the other modality, whenever one of them is noisy, too weak or totally missing. Using the OSPA-T and MOT scores we showed that the proposed method outperforms the PF-based method [4].

Appendix A
An Audio Generative Model

In this appendix we describe the audio observation model used in this paper. More precisely, we make explicit the

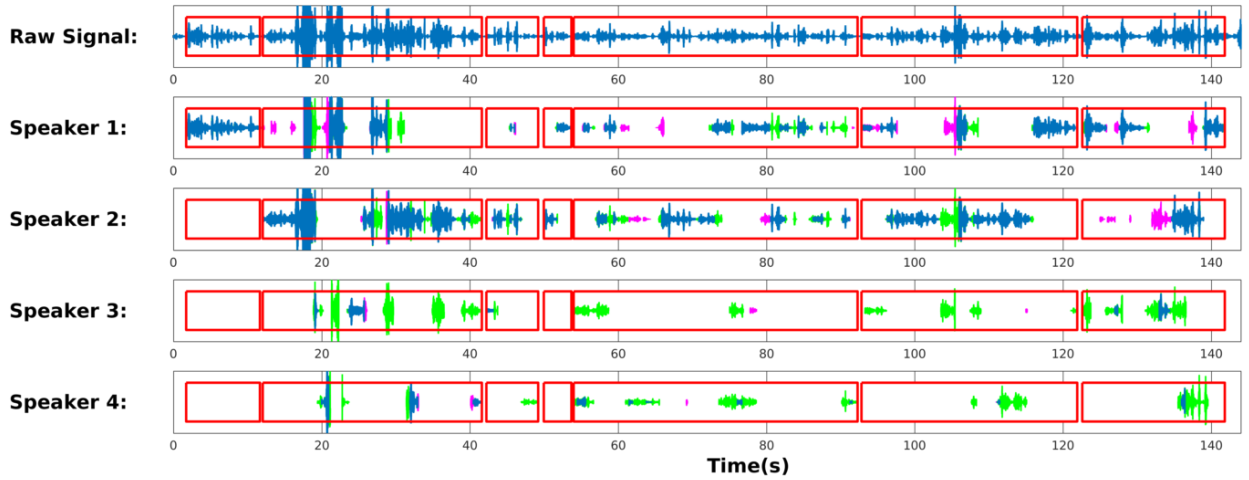


Fig. 7: Diarization results obtained with Seq13-4P-S2M1 (FFOV). The first row shows the audio signal recorded with one of the microphones. The red boxes show the result of the voice activity detector which is applied to all the microphone signals prior to tracking. For each speaker, correct detections are shown in blue, missed detections are shown in green, and false positives are shown in magenta

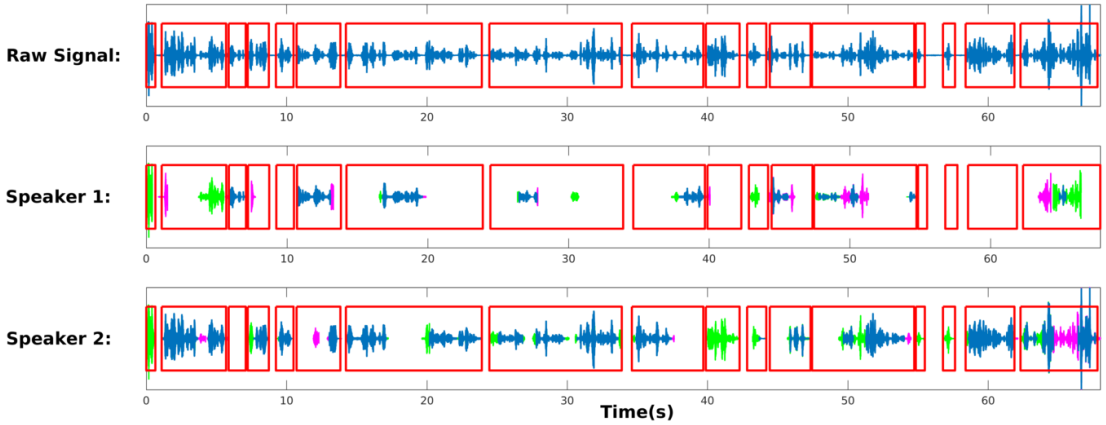


Fig. 8: Diarization results obtained with Seq19-2P-S1M1 (PFOV).

generative model introduced in Section III-D, i.e. equation (13). For that purpose we consider a training set of audio features, or inter-channel spectral features (which in practice correspond to the real and imaginary parts of complex-valued Fourier coefficients) and their associated source locations, $\mathcal{T} = \{(\boldsymbol{g}_i, \boldsymbol{x}_i)\}_{i=1}^{I}$ and let $(\boldsymbol{g}, \boldsymbol{x}) \in \mathcal{T}$. The vector $\boldsymbol{g}$ is the concatenation of K vectors $\boldsymbol{g} = [\boldsymbol{g}_1|\dots\boldsymbol{g}_k|\dots\boldsymbol{g}_K]$ where $[\cdot|\cdot]$ denotes vertical vector concatenation. We recall that for all sub-bands $k; 1 \leq k \leq K$, $\boldsymbol{g}_k \in \mathbb{R}^{2J}$ where J is the number of frequencies in each sub-band. Without loss of generality we consider the sub-band k. The joint probability of $(\boldsymbol{g}_k, \boldsymbol{x})$ can be marginalized as:

$$p(\boldsymbol{g}_k, \boldsymbol{x}) = \sum_{r=1}^{R} p(\boldsymbol{g}_k|\boldsymbol{x}, C_k = r)p(\boldsymbol{x}_k|C_k = r)p(C_k = r). \tag{36}$$

Assuming Gaussian variables, we have $p(\boldsymbol{g}_k|\boldsymbol{x}, C_k = r) = \mathcal{N}(\boldsymbol{g}_k|h_{kr}(\boldsymbol{x}), \boldsymbol{\Sigma}_{kr})$, $p(\boldsymbol{x}|C_k = r) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\nu}_{kr}, \boldsymbol{\Omega}_{kr})$, and $p(C_k = r) = \pi_{kr}$, where $h_{kr}(\boldsymbol{x}) = \mathbf{L}_{kr}\boldsymbol{x} + \boldsymbol{l}_{kr}$ with $\mathbf{L}_{kr} \in \mathbb{R}^{2J\times 2}$ and $\boldsymbol{l}_{kr} \in \mathbb{R}^{2J}$, $\boldsymbol{\Sigma}_r \in \mathbb{R}^{2J\times 2J}$ is the associated covariance matrix, and $\boldsymbol{x}$ is drawn from a Gaussian mixture model with R components, each component r being characterized by a prior π_{kr}, a mean $\boldsymbol{\nu}_{kr} \in \mathbb{R}^2$ and a covariance $\boldsymbol{\Omega}_{kr} \in \mathbb{R}^{2\times 2}$. The parameter set of this model for sub-band k is:

$$\boldsymbol{\Theta}_k = \{\mathbf{L}_{kr}, \boldsymbol{l}_{kr}, \boldsymbol{\Sigma}_{kr}, \boldsymbol{\nu}_{kr}, \boldsymbol{\Omega}_{kr}, \pi_{kr}\}_{r=1}^{r=R}. \tag{37}$$

These parameters can be estimated via a closed-form EM procedure from a training dataset, e.g. $\mathcal{T}$ (please consult [15], [29] and Section VII-C).

One should notice that there is a parameter set for each sub-band k, $1 \leq k \leq K$, hence there are K models that need be trained in our case. It follows that (12) writes:

$$p(\boldsymbol{g}_{tk}|\boldsymbol{x}_{tn}, B_{tk} = n, C_{tk} = r) = \begin{cases} \mathcal{N}(\boldsymbol{g}_{tk}; \mathbf{L}_{kr}\boldsymbol{x}_{tn} + \boldsymbol{l}_{kr}, \boldsymbol{\Sigma}_{kr}) & \text{if } 1 \leq n \leq N \\ \mathcal{U}(\boldsymbol{g}_{tk}; \text{vol}(\mathcal{G})) & \text{if } n = 0. \end{cases} \tag{38}$$

The right-hand side of (7) can now be written as:

$$p(C_{tk} = r|\boldsymbol{x}_{tn}, B_{tk} = n) = \frac{\pi_r \mathcal{N}(\boldsymbol{x}_{tn}; \boldsymbol{\nu}_r, \boldsymbol{\Omega}_r)}{\sum_{i=1}^{R} \pi_i \mathcal{N}(\boldsymbol{x}_{tn}; \boldsymbol{\nu}_i, \boldsymbol{\Omega}_i)}. \tag{39}$$

APPENDIX B
DERIVATION OF THE E-S VARIATIONAL STEP

The E-S step for the per-person variational posterior distribution of the state vector $q(\boldsymbol{s}_{tn})$ is evaluated by expanding (16), namely:

$$\begin{aligned}
\log\, q(\boldsymbol{s}_{tn}) &= \mathbb{E}_{q(\boldsymbol{z}_t)\prod_{\ell\neq n} q(\boldsymbol{s}_{t\ell})}[\log p(\mathbf{f}_t|\mathbf{s}_t,\boldsymbol{a}_t)p(\mathbf{g}_t|\mathbf{s}_t,\boldsymbol{b}_t,\boldsymbol{c}_t)\\
&\times p(\boldsymbol{z}_t|\mathbf{s}_t)\int p(\mathbf{s}_t|\mathbf{s}_{t-1})q(\mathbf{s}_{t-1}|\mathbf{o}_{1:t-1})d\mathbf{s}_{t-1}]\\
&= \sum_{m=1}^{M_t}\mathbb{E}_{q(\boldsymbol{A}_t)}[\delta_{(\boldsymbol{A}_{tm}=n)}\log\mathcal{N}(\boldsymbol{v}_{tm};\mathbf{P}_f\boldsymbol{s}_{tn},\boldsymbol{\Phi}_{tm})]\\
&+ \sum_{k=1}^{K_t}\mathbb{E}_{q(\boldsymbol{B}_t,\boldsymbol{C}_t)}[\delta_{(\boldsymbol{B}_{tk}=n,\boldsymbol{C}_{tk}=r)}\log\mathcal{N}(\boldsymbol{g}_{tk};h_{kr}(\boldsymbol{x}_{tn}),\boldsymbol{\Sigma}_{kr})]\\
&+ \mathbb{E}_{\prod_{\ell\neq n} q(\boldsymbol{s}_{t\ell})}[\sum_{n=1}^{N}\log\mathcal{N}(\boldsymbol{s}_{tn};\mathbf{D}\boldsymbol{s}_{t-1n},\mathbf{D}\Gamma_{t-1n}\mathbf{D}^\top+\Lambda_{tn})]
\end{aligned}$$

where constant terms are omitted. Using (20) and after some algebraic derivations one obtains that $q(\boldsymbol{s}_{tn})$ follows a Gaussian distribution, i.e. (22), where the covariance matrix and mean vector are given by (23) and (24), respectively.

REFERENCES

[1] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Audiovisual probabilistic tracking of multiple speakers in meetings," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 601–616, 2007.

[2] T. Hospedales and S. Vijayakumar, "Structure inference for Bayesian multisensory scene understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2140–2157, 2008.

[3] S. Naqvi, M. Yu, and J. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 895 –910, 2010.

[4] V. Kılıç, M. Barnard, W. Wang, and J. Kittler, "Audio assisted robust visual tracking with adaptive particle filtering," *IEEE Transactions on Multimedia*, vol. 17, no. 2, pp. 186–200, 2015.

[5] N. Schult, T. Reineking, T. Kluss, and C. Zetzsche, "Information-driven active audio-visual source localization," *PloS one*, vol. 10, no. 9, 2015.

[6] M. Barnard, W. Wang, A. Hilton, J. Kittler *et al.*, "Mean-shift and sparse sampling-based SMC-PHD filtering for audio informed visual speaker tracking," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2417–2431, 2016.

[7] V. Kılıç, M. Barnard, W. Wang, A. Hilton, and J. Kittler, "Mean-shift and sparse sampling-based SMC-PHD filtering for audio informed visual speaker tracking," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2417–2431, 2016.

[8] S. Ba, X. Alameda-Pineda, A. Xompero, and R. Horaud, "An on-line variational Bayesian model for multi-person tracking from cluttered scenes," *Computer Vision and Image Understanding*, vol. 153, pp. 64–76, 2016.

[9] S. Bae and K. Yoon, "Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 595–610, March 2018.

[10] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, no. 3, pp. 216–228, 2007.

[11] A. Lombard, Y. Zheng, H. Buchner, and W. Kellermann, "TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1490–1503, 2011.

[12] X. Alameda-Pineda and R. Horaud, "A geometric approach to sound source localization from time-delay estimates," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 6, pp. 1082–1095, 2014.

[13] Y. Dorfan and S. Gannot, "Tree-based recursive expectation-maximization algorithm for localization of acoustic sources," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 10, pp. 1692–1703, 2015.

[14] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2171–2186, 2016.

[15] X. Li, L. Girin, R. Horaud, S. Gannot, X. Li, L. Girin, R. Horaud, and S. Gannot, "Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 10, pp. 1997–2012, 2017.

[16] A. Deleforge, R. Horaud, Y. Y. Schechner, and L. Girin, "Co-localization of audio sources in images using binaural features and locally-linear regression," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 4, pp. 718–731, 2015.

[17] B. Gold, N. Morgan, and D. Ellis, *Speech and audio signal processing: processing and perception of speech and music*. John Wiley & Sons, 2011.

[18] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: An audio-visual corpus for speaker localization and tracking," in *Machine Learning for Multimodal Interaction*. Springer, 2004, pp. 182–195.

[19] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal Bayesian fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1086–1099, 2018.

[20] B. Ristic, B.-N. Vo, D. Clark, and B.-T. Vo, "A metric for performance evaluation of multi-target tracking algorithms," *IEEE Transactions on Signal Processing*, vol. 59, no. 7, pp. 3452–3457, 2011.

[21] D. Vijayasenan and F. Valente, "DiarTk: an open source toolkit for research in multistream speaker diarization and its application to meeting recordings," in *INTERSPEECH*, Portland, OR, USA, 2012, pp. 2170–2173.

[22] N. Checka, K. Wilson, M. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, pp. 881–884.

[23] Y. Liu, W. Wang, J. Chambers, V. Kilic, and A. Hilton, "Particle flow SMC-PHD filter for audio-visual multi-speaker tracking," in *International Conference on Latent Variable Analysis and Signal Separation*, 2017, pp. 344–353.

[24] Y. Liu, A. Hilton, J. Chambers, Y. Zhao, and W. Wang, "Non-zero diffusion particle flow SMC-PHD filter for audio-visual multi-speaker tracking," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 4304–4308, April 2018.

[25] X. Qian, A. Brutti, M. Omologo, and A. Cavallaro, "3D audio-visual speaker tracking with an adaptive particle filter," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, New-Orleans, Louisiana, 2017, pp. 2896–2900.

[26] I. D. Gebru, X. Alameda-Pineda, F. Forbes, and R. Horaud, "EM algorithms for weighted-data clustering with application to audio-visual scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 12, pp. 2402–2415, 2016.

[27] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, "Online Localization and Tracking of Multiple Moving Speakers in Reverberant Environments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 88–103, Mar. 2019.

[28] Y. Ban, X. Alameda-Pineda, C. Evers, and R. Horaud, "Tracking Multiple Audio Sources with the Von Mises Distribution and Variational EM," *IEEE Signal Processing Letters*, vol. 26, no. 6, pp. 798 – 802, Jun. 2019, paper submitted to IEEE Signal Processing Letters.

[29] Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, "Exploiting the complementarity of audio and visual data in multi-speaker tracking," in *IEEE ICCV Workshop on Computer Vision for Audio-Visual Media*, Venezia, Italy, Oct. 2017, pp. 446–454.

[30] Y. Ban, X. Li, X. Alameda-Pineda, L. Girin, and R. Horaud, "Accounting for room acoustics in audio-visual multi-speaker tracking," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, Alberta, Canada, Apr. 2018, pp. 6553–6557.

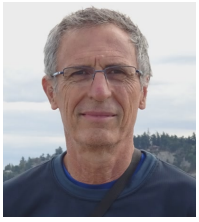
[31] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, pp. 99–109, 1943.

[32] A. Deleforge, F. Forbes, and R. Horaud, "High-dimensional regression with Gaussian mixtures and partially-latent response variables," *Statistics and Computing*, vol. 25, no. 5, pp. 893–911, 2015.
[33] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
[34] V. Smidl and A. Quinn, *The Variational Bayes Method in Signal Processing*. Springer, 2006.
[35] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.
[36] A. Noulas, G. Englebienne, and B. J. A. Krose, "Multimodal speaker diarization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 79–93, 2012.
[37] G. Lathoud and M. Magimai-Doss, "A sector-based, frequency-domain approach to detection and localization of multiple speakers," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3. IEEE, 2005, pp. 265–268.
[38] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii, USA, 2017, pp. 7291–7299.
[39] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii, USA, 2017, pp. 1367–1376.
[40] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831*, 2016.
[41] V. P. Minotto, C. R. Jung, and B. Lee, "Multimodal multi-channel on-line speaker diarization using sensor fusion through SVM," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1694–1705, 2015.

Yutong Ban received the B.Sc degree in telecommunications from Xidian University, China and the engineering degree in image processing and computer science from Télécom Saint-Etienne, France. He is currently a Ph.D student in the PERCEPTION team at Inria Grenoble. His research interests include audio-visual fusion, speaker diarization and multi-object tracking.

Xavier Alameda-Pineda received M.Sc. degrees in mathematics (2008), in telecommunications (2009) and in computer science (2010) and a Ph.D. in mathematics and computer science (2013) from Université Joseph Fourier. Since 2016, he is a research scientist at Inria Grenoble Rhône-Alpes, with the Perception team. He served as Area Chair at ICCV'17, of ICIAP'19 and of ACM MM'19. He is the recipient of several paper awards and of the ACM SIGMM Rising Star Award in 2018. His scientific interests lie in computer vision, machine learning and signal processing for robotics and multimodal social behavior analysis.

Laurent Girin received the M.Sc. (1994) and Ph.D. (1997) degrees in signal processing from Institut National Polytechnique de Grenoble (INPG), France. In 1999 he joined Ecole Nationale Supérieure d'Electronique et de Radioélectricité de Grenoble, as an associate professor. Currently his is a professor at Physics, Electronics, and Materials Department of INPG, where he lectures signal processing theory and applications to audio. His research activity is carried out at GIPSA-Lab (Grenoble Laboratory of Image, Speech, Signal, and Automation). It deals with speech and audio processing (analysis, modeling, coding, transformation, synthesis), with a special interest in multimodal speech processing (e.g. audiovisual, articulatory-acoustic, etc.) and speech/audio source separation. Prof. Girin regularly collaborates with the PERCEPTION team at Inria Grenoble.



Radu Horaud received the B.Sc. degree in Electrical Engineering, the M.Sc. degree in Control Engineering, and the Ph.D. degree in Computer Science from the Institut National Polytechnique de Grenoble, France. In the past he held research positions with SRI International (1982-1984) and with CNRS (1984-1998). Since 1998 he has been director of research (the equivalent of full processor) with Inria Grenoble Rhône-Alpes. He is the founder and head of the PERCEPTION team. His research interests include computer vision, machine learning, audio signal processing, audio-visual analysis, and robotics. Radu Horaud and his collaborators received numerous best paper awards. He was an area editor of the *Elsevier Computer Vision and Image Understanding* (1999-2017), he is a member of the advisory board of the *Sage International Journal of Robotics Research* and an associate editor of the *Kluwer International Journal of Computer Vision*. He was program co-chair of IEEE ICCV'01 and of ACM ICMI'15. Radu Horaud was awarded two ERC grants, an advanced grant for the project *Vision and Hearing in Action* (2014-2019) and a proof of concept grant (2018-2019).